

A PARTIALLY-AUTOMATED APPROACH TO
THE ASSESSMENT OF MATHEMATICS IN
HIGHER EDUCATION

PETER JAMES ROWLETT

A thesis submitted in partial fulfilment of the requirements of
Nottingham Trent University for the degree of Doctor of Philosophy

July 2013

This work is the intellectual property of the author. You may copy up to 5% of this work for private study, or personal, non-commercial research. Any re-use of the information contained within this document should be fully referenced, quoting the author, title, university, degree level and pagination. Queries or requests for any other use, or if a more substantial copy is required, should be directed in the owner(s) of the Intellectual Property Rights.

This thesis was typeset in Bitstream Vera Serif, Gothic Uralic and `Nimbus Mono L` in LibreOffice 3.6 on Ubuntu 12.10. Statistical analysis was conducted and plots were produced using R 2.15.1.

Abstract

E-assessment in higher education mathematics is explored via a systematic review of literature and a practitioner survey, and compared with other assessment approaches in common use in higher education mathematics in the UK. E-assessment offers certain advantages over other approaches, for example question randomisation allows individualisation of assessment, but it is restricted in the range of what can be assessed due to the limitations of automated marking.

A partially-automated approach is proposed in which e-assessment techniques are used to set an individualised assessment which is taken and marked by hand. This approach is implemented in a higher education mathematics module. The module uses individual coursework assignments alongside group work to attempt to account for individual contribution to learning outcomes. The partially-automated approach is used as a method for reducing the risk of plagiarism in this coursework, rather than replacing it with a written examination or e-assessment.

Evaluation via blind second-marking indicates that the approach was capable of setting a reliable and valid assessment. Evaluation of student views and analysis of assessment marks leads to the conclusion that plagiarism does take place among the undergraduate cohort, was a risk during this assessment, but was not in fact a particular problem.

The partially-automated approach is recommended as an appropriate addition to the repertoire of higher education mathematics assessment methods, particularly in cases where an assessment carries a high risk of plagiarism but the need for open-ended or deeper questions make an examination or automated marking system sub-optimal.

Acknowledgements

This project has been a long journey, and consequently there are many people to thank.

For their direct involvement with the work reported, I thank: four pseudonymised interviewees and 38 anonymous questionnaire respondents who contributed the data for chapter 3; seven anonymous respondents to the pilot questionnaire for that survey; my students who took part in the teaching described in sections 2.2, 5.3 and 7.2 and provided feedback; one unnamed lecturer and his students who took part in the experiment described in section 7.3.2 and provided feedback; and, three pseudonymised second-marker volunteers and two unnamed lecturers who provided work for me to second mark for the evaluation described in section 7.3.1. I am grateful to Christian Perfect, School of Mathematics and Statistics, Newcastle University, for adapting the Numbas e-assessment system for my main experiment, described principally in section 7.2.

I thank Gizem Karaali, Mark Huber and the anonymous reviewers at the *Journal of Humanistic Mathematics* for comments and encouragement which led me to improve section 2.2. In addition, for publishing work relevant to this thesis, I thank: Michael Grove, Janet Nuttall, Robert Surowiec and Chantal Jackson at *MSOR Connections*; editors and administrators of the proceedings for the 2010 and 2011 NTU Annual Learning and Teaching Conferences; and, Jeff Waldock.

Apart from the organisers of various conferences who accepted my contribution to their programmes, for inviting me to speak about work included in this thesis I thank: Stephen Hibberd; Jeff Waldock; Bill Foster and Christian Perfect; Thomas Prellberg and Francis Wright; and, Stevie Vanhegan. I also thank attendees at the presentations I have given, whose reactions, comments and questions helped shape this work.

This project was part-time and self-funded. I thank those who employed me along the way and provided useful experiences which fed into its development, including Neville Davies, David Youdan, Oliver Jensen, Duncan Lawson, Graham Rogerson and Christian Beardah.

Otherwise, for specific help or advice on work that is included in this thesis, I thank Sally Barton, Christian Beardah, David Bedford, Jeanette Bisschop, Blogger users 'kensson' and 'Xittenn', Noel-Ann Bradshaw, Pete Bradshaw, Graham Brindley, Doreen Connor, Barrie Cooper, Stephen Cox, Neville Davies, Pete Dixon, Rob Dransfield, David Fairhurst, Joel Feinstein, Mike Frost, Samuel Hansen, Stephen Hibberd, Neville Hunt, Sharon Hutchings, Marina Isaac, Tony Mann, Stephen O'Brien, Marek Oledzki, Martyn Parker, Brad Payne, Christian Perfect, John Read, Daniel Rhodes, Calvin Smith, Katie Steckles, Kathryn Taylor, Jonathan Townsend, Jeff Waldock, Dagmar Waller and Sarah Woodhouse.

I thank my supervisors, Graham Tranfield (including as Director of Studies 2010-13), Tony Sackfield (Director of Studies 2004-9) and Wayne Cranton, for looking after me and all their help and advice during this project. I thank my internal assessors John Gough (2004-6), Mike Baxter (2007-9) and James Hind (2010-13), who kept an eye on the project as it developed and offered useful advice. I thank my examiners Neil Challis and Jonathan Tepper, for their time and effort.

I thank my parents, Sue and Richard, for their support and for never giving me the idea that I couldn't do it.

For help in innumerable ways during this long stretch, I thank my wife Emma.

Summary table of contents

Abstract.....	ii
Acknowledgements.....	iii
Detailed table of contents.....	vi
Index of tables.....	xiii
Index of figures.....	xv
1 Introduction.....	1
2 Background.....	8
3 Survey of mathematics lecturers' views of e-assessment.....	50
4 Discussion of assessment methods and proposal of a novel approach.....	64
5 Teaching and learning context.....	76
6 Evaluation method.....	94
7 Implementation.....	111
8 Results.....	142
9 Conclusions and discussion.....	163
References.....	185
Appendices.....	202

Detailed table of contents

Abstract.....	ii
Acknowledgements.....	iii
Index of tables.....	xiii
Index of figures.....	xv
1 Introduction.....	1
1.1 Outline of research.....	1
1.2 Research questions.....	4
1.2.1 Main research questions.....	4
1.2.2 Questions related to the particular implementation and its evaluation.....	4
1.3 Original contribution.....	5
1.4 Overview of chapters.....	5
2 Background.....	8
2.1 Introduction.....	8
2.2 Reflection on experience.....	9
2.2.1 Introduction.....	9
2.2.2 A solution looking for a problem.....	10
2.2.3 Deep learning of mathematical concepts.....	11
2.2.4 Development of graduate skills	13
2.2.5 The pattern of the technology enthusiast.....	15
2.2.6 Discussion.....	17
2.3 Assessment in higher education mathematics.....	19
2.3.1 Introduction.....	19
2.3.2 About assessment and its evaluation.....	20
2.3.2.1 What is assessment?.....	20
2.3.2.2 The purpose of a mathematics degree.....	21
2.3.2.3 Purposes of assessment	22
2.3.2.4 Assessment of graduate skills.....	23
2.3.2.5 Learning outcomes and alignment.....	25
2.3.2.6 Formative and summative assessment.....	25

2.3.2.7 Marking criteria.....	25
2.3.2.8 Plagiarism.....	26
2.3.2.9 Evaluation of assessment.....	27
2.3.3 Methods of assessment used.....	30
2.3.3.1 Introduction.....	30
2.3.3.2 Written examination.....	31
2.3.3.3 Coursework.....	33
2.3.3.4 Presentations.....	34
2.3.3.5 Peer assessment.....	35
2.3.3.6 Posters.....	35
2.3.3.7 Oral assessment.....	35
2.3.3.8 A note about group work.....	36
2.4 Systematic review of literature on advantages and limitations of e-assessment	38
2.4.1 Introduction.....	38
2.4.2 Method.....	38
2.4.3 Findings overview.....	39
2.4.4 Types of e-assessment and their uses.....	40
2.4.5 Advantages of e-assessment.....	45
2.4.6 Limitations of e-assessment.....	47
3 Survey of mathematics lecturers' views of e-assessment.....	50
3.1 Introduction.....	50
3.2 Method.....	51
3.3 Background.....	52
3.4 Data.....	54
3.4.1 Suitability of traditional and e-assessment.....	54
3.4.2 Advantages and limitations of e-assessment	55
3.4.3 Producing mathematics.....	58
3.4.4 Marking.....	59
3.4.5 Own use of computers.....	60
3.5 Different groupings.....	61
3.6 Findings of the literature review and lecturer survey on e-assessment.....	62

4 Discussion of assessment methods and proposal of a novel approach.....	64
4.1 Introduction.....	64
4.2 Assessment methods in common use.....	64
4.3 Desirable qualities of assessment.....	65
4.4 Discussion of assessment methods in terms of the desirable qualities.....	66
4.4.1 Introduction.....	66
4.4.2 Written examination.....	67
4.4.3 Frequent short coursework.....	68
4.4.4 Longer coursework assignments	68
4.4.5 Presentations.....	69
4.4.6 Peer assessment	70
4.4.7 Posters.....	70
4.4.8 Oral assessment.....	70
4.4.9 E-assessment.....	71
4.5 Proposal for a partially-automated approach to the assessment of mathematics in higher education.....	72
5 Teaching and learning context.....	76
5.1 Introduction.....	76
5.2 Programme and module context.....	76
5.2.1 University graduate attributes.....	76
5.2.2 Degree programme context.....	77
5.2.3 Module aims and objectives.....	78
5.3 Previous experience of group work.....	79
5.3.1 Teaching and learning context.....	79
5.3.2 Module design.....	80
5.3.3 The topic.....	81
5.3.4 The project	84
5.3.5 Evaluation of previous group project activity.....	86
5.3.5.1 Method.....	86
5.3.5.2 Suitability of Art Gallery Problems as a topic	86
5.3.5.3 Measures to tackle uneven contribution.....	87
5.3.6 Discussion.....	89

5.4 Group projects.....	90
5.5 Group management, individual work and uneven contribution.....	91
6 Evaluation method.....	94
6.1 Introduction.....	94
6.2 What will the evaluation do?.....	95
6.3 Who is the evaluation for?	96
6.4 Practicalities.....	97
6.5 Methods.....	98
6.5.1 Are the marks particularly sensitive to who is doing the marking?.....	99
6.5.2 Is the assignment assessing the learning outcomes it was intended to assess?	100
6.5.3 Does the individualised nature of the assignment work to reduce plagiarism?.....	101
6.5.4 Can the individual work contribute to a view on whether uneven contribution is taking place?.....	101
6.5.5 Can the individual work contribute to a process for adjusting group marks to take account of individual contribution?.....	102
6.5.6 What are student views on conventional measures to address uneven contribution?.....	103
6.5.7 What are student views on the purpose of assessment?.....	103
6.5.8 What are student views on automated marking?.....	103
6.6 Potential negative impact and ethics.....	104
6.7 Deliverables and plans for positive impact.....	105
6.8 Concepts and tests.....	106
6.8.1 Properties of assessment marks.....	106
6.8.2 Marks standardisation.....	108
6.8.3 Measures of correlation.....	108
6.8.3.1 Tukey mean-difference plot.....	108
6.8.3.2 Pearson's product-moment correlation coefficient (Pearson's ρ).....	108
6.8.3.3 Kendall's rank correlation coefficient (Kendall's τ).....	109
6.8.4 Inter-rater reliability.....	109
6.8.5 Comparison of interval or ordinal data from two groups.....	110

7 Implementation.....	111
7.1 Introduction.....	111
7.2 Individual assignment 2.....	112
7.2.1 Technology.....	112
7.2.1.1 Introduction.....	112
7.2.1.2 Design.....	113
7.2.1.3 Prototyping.....	114
7.2.1.4 Adoption of Numbas.....	125
7.2.2 Implementation of individual assignment 2 in Numbas.....	127
7.3 Evaluation.....	134
7.3.1 Second-marker experiment.....	134
7.3.1.1 Method.....	134
7.3.1.2 Written examination reference experiment.....	135
7.3.1.3 Coursework reference experiment.....	135
7.3.1.4 Second-marking of individual assignment 2.....	135
7.3.2 Student feedback questionnaire.....	137
7.3.2.1 Method.....	137
7.3.2.2 Student views on conventional measures to address uneven contribution.....	137
7.3.2.3 Student views on individualised assessment.....	138
7.3.3 Comparison of marks.....	139
7.3.3.1 Measure of evenness of contribution to group marks.....	139
7.3.3.2 Correlation between individual and group marks.....	140
7.3.3.3 Comparison with peer assessment of contribution.....	140
8 Results.....	142
8.1 Second-marking.....	142
8.1.1 Written examination reference experiment.....	142
8.1.2 Coursework reference experiment.....	143
8.1.3 Second-marking of individual assignment 2.....	145
8.1.3.1 Evaluation of marker consistency.....	145
8.1.3.2 Comments on learning outcomes.....	147
8.1.3.3 Comments on process.....	148

8.2 Student feedback.....	148
8.2.1 Student views on conventional measures to address uneven contribution	148
8.2.1.1 Minutes of meetings.....	148
8.2.1.2 Peer assessment of contribution.....	149
8.2.2 Student views on individualised assessment.....	151
8.2.2.1 Students in this group project.....	151
8.2.2.2 Reference group of students at another university.....	154
8.2.2.3 Comparison between the two groups.....	156
8.3 Marks for individual assignment 2.....	158
8.3.1 Measure of evenness of contribution to group marks.....	158
8.3.2 Correlation between individual and group marks.....	159
8.3.2.1 Reference experiment.....	159
8.3.2.2 Correlation between group mean individual marks and marks for group work.....	159
8.3.3 Comparison with peer assessment of contribution.....	160
9 Conclusions and discussion.....	163
9.1 Introduction.....	163
9.2 An examination of e-assessment in higher education mathematics assessment	163
9.2.1 Literature review and survey.....	164
9.2.2 Student feedback.....	165
9.3 Is there an opportunity for a novel approach?.....	166
9.4 Is there a learning and teaching context in which the proposed novel approach is more effective than existing methods?.....	168
9.4.1 Are the marks particularly sensitive to who is doing the marking?.....	169
9.4.2 Is the assignment assessing the learning outcomes it was intended to assess?.....	171
9.4.3 Does the individualised nature of the assignment work to reduce plagiarism?.....	171
9.4.4 Can the individual work contribute to a view on whether uneven contribution is taking place?.....	172

9.4.5 Can the individual work contribute to a process for adjusting group marks to take account of individual contribution?.....	173
9.4.6 What are student views on conventional measures to address uneven contribution?.....	174
9.5 Final conclusions for this project.....	174
9.6 Limitations of this study and opportunities for further work.....	176
9.6.1 The partially-automated approach.....	176
9.6.2 E-assessment.....	178
9.6.3 Student feedback survey.....	179
9.6.4 Second-marker experiments.....	179
9.6.5 Uneven contribution.....	181
9.6.6 Future practice.....	183
References.....	185
Appendices.....	202
A. List of relevant publications.....	202
B. List of relevant presentations.....	204
Seminars and conference talks.....	204
Non-live presentations.....	207
C. Information given to second-markers of individual assignment 2.....	208
Grade descriptions (marking-notes.pdf).....	208
Mark scheme (marking-scheme.pdf).....	209
Sample piece of marked work.....	212
Question sheet.....	212
Answer sheet.....	212
Sample answer.....	213

Index of tables

Table 3.1: Information about interview participants.....	53
Table 3.2: Questionnaire responses to: 'Please indicate which of the following applies to you'.....	53
Table 3.3: Questionnaire responses to: 'Do you consider yourself to be a user of e-assessment in your teaching?'.....	54
Table 3.4: How suitable questionnaire respondents felt traditional and e-assessment are for four areas of assessment.....	55
Table 3.5: Questionnaire respondents' levels of agreement with 12 statements.....	58
Table 3.6: Questionnaire respondents' 'yes/no' responses to four statements about their use of computers in setting and marking work.....	61
Table 5.1: Free text examples of previous knowledge or skills required for this project.....	87
Table 5.2: Responses to 'How would you rate the difficulty of the mathematical content of this project compared with your other modules?'.....	87
Table 5.3: Student responses to two questions about the plan of work and minutes of meetings.....	88
Table 5.4: Student responses to 'All team members contributed to the project equally'.	88
Table 5.5: Breakdown of responses to 'How helpful did you feel the minutes of meetings were in ensuring team members completed their assigned tasks?' by team role.....	89
Table 8.1: Original and second marking for ten written examinations.....	142
Table 8.2: Original and second marking for fourteen pieces of open-ended coursework.....	144
Table 8.3: Original and second marks for five pieces of work submitted for individual assignment 2.....	145
Table 8.4: Student responses to 'How helpful did you feel the minutes of meetings were in ensuring team members completed their assigned tasks?'.....	149
Table 8.5: Number of students disagreeing and agreeing for four statements.....	152

Table 8.6: Student choices for “What are the most important aspects of assessment?”	152
Table 8.7: Number of students disagreeing and agreeing for four questions.....	155
Table 8.8: Student choices for “What are the most important aspects of assessment?”	156
Table 8.9: p-values for Fisher's Exact Test on the four Likert scale questions between the two groups of students.....	157
Table 8.10: Marks range for individual assignment marks within each group.....	158
Table 8.11: Standard deviation for individual assignment marks within each group.	158
Table 8.12: Means for individual assignment marks within each group.....	159
Table 8.13: Raw group marks for group project 2.....	159
Table 8.14: Values of Kendall's correlation coefficient for different definitions of the range of even contribution.....	162

Index of figures

Figure 5.1: Convex polygon guarded by one guard.....	82
Figure 5.2: L-shaped gallery with a single guard.....	82
Figure 5.3: Gallery which requires more than one guard.....	82
Figure 5.4: Gallery which requires $\left\lceil \frac{n}{3} \right\rceil$ guards, where n is the number of vertices.....	83
Figure 5.5: Triangulation and vertex colouring of gallery from figure 5.3.....	84
Figure 5.6: Triangulation and vertex colouring of gallery from figure 5.4.....	84
Figure 6.1: Histogram of marks for individual assignment 2.....	107
Figure 7.1: Inserting dollar-delimited LaTeX code in a TinyMCE interface.....	115
Figure 7.2: The preview stage echoes LaTeX back for LaTeXMathML to interpret...	116
Figure 7.3: The stored MathML code is presented as published content.....	116
Figure 7.4: The variables are echoed back for the user to select which will be pseudo-randomised.....	117
Figure 7.5: Input via dollar-delimited LaTeX in TinyMCE using @ identifiers for pseudo-randomised constants.....	117
Figure 7.6: Randomisation options for identified constants.....	118
Figure 7.7: Worksheet template interface. The question is marked in blue, and the answer in red.....	120
Figure 7.8: Content block input interface showing text input.....	121
Figure 7.9: Content block input interface showing LaTeX input.....	122
Figure 7.10: Question sheet input.....	122
Figure 7.11: Answer sheet input.....	123
Figure 7.12: Interpreted question sheet.....	123
Figure 7.13: Interpreted answer sheet.....	124
Figure 7.14: Generated question sheets.....	124
Figure 7.15: Generated answer sheets.....	125
Figure 7.16: Nine museum floor plans, each requiring 3, 4 or 5 guards.....	129
Figure 7.17: Numbas question input interface showing question 1.....	131
Figure 7.18: Numbas answer input showing question 1.....	131
Figure 7.19: Numbas question 'variable' definition showing variables for question 1.	

.....	132
Figure 7.20: Numbas-generated question sheet and worksheet generation controls.	
.....	133
Figure 7.21: Numbas-generated answer sheet.....	133
Figure 8.1: Standardised Tukey mean-difference plot for the written examination reference experiment.....	143
Figure 8.2: Standardised Tukey mean-difference plot for the coursework reference experiment.....	144
Figure 8.3: Standardised Tukey mean-difference plot with second-marker A for individual assignment 2.....	145
Figure 8.4: Standardised Tukey mean-difference plot with second-marker B for individual assignment 2.....	146
Figure 8.5: Standardised Tukey mean-difference plot with second-marker C for individual assignment 2.....	146
Figure 8.6: Plot of values of Kendall's correlation coefficient for different definitions of the range of even contribution.....	162

1 Introduction

1.1 Outline of research

This research examines the use of 'e-assessment' (assessment set and marked by computer; sometimes called computer-aided or computer-assisted assessment, CAA) in higher education mathematics, contrasted with 'traditional' assessment approaches (set and marked through non-automated methods). An investigation of assessment approaches leads to the proposal of a novel, partially-automated approach, which I implement and evaluate in a particular teaching and learning context.

This research originated in an earlier project, which drew as its motivation an enthusiastic desire to implement e-assessment technology simply because implementation was possible, without particular study of educational need or purpose. My professional development as a teacher in higher education resulted in a change from this naïve approach to a more sceptical one, where innovation follows educational need and is only completed where it is more advantageous than existing approaches.

Moore (2011) predicts that “your focus will shift through the life-cycle of [a] project as it evolves and takes shape” (p. 6). Indeed, this is a thesis of three parts, representing three phases of this research.

The first part (chapters 2-3) studies assessment in higher education mathematics in general, and e-assessment in higher education mathematics in particular. This is a free, open and unfocused exploration of the topic, given that “the more certain you are about something in education, the less likely you are to appreciate what is going

on" (Mason, 2002; p. 120). As the approach taken is to implement an innovation only where there exists some educational need, the aim of this investigation is to study the place of technology in assessment in higher education mathematics in order to identify an opportunity for novel development. A weakness is identified in the literature, in that implementation of e-assessment, particularly that reported in published work, is often driven by individual enthusiasts (Barton et al, 2012; p. 18), which may exaggerate the benefits (Sangwin, 2012; p. 12). A survey is thus undertaken with a focus on users of e-assessment who are not represented in the literature and non-users of e-assessment.

The second part (chapters 4-5) discusses different methods of assessment in terms of their advantages and limitations. It is recognised that different assessment methods in different teaching and learning contexts access certain desirable qualities of assessment to different extents. For example, a written examination can be more reliable than a piece of open-ended coursework, but may be less valid, in terms of being able to assess everything a mathematics degree should contain. Discussion of each assessment method as a balance of desirable qualities leads to the identification of a partially-automated approach, in which the tools of e-assessment are used to set an individualised (pseudo-randomised) assessment which may be taken and marked offline. This approach is claimed as novel in higher education mathematics, since a similar approach was not found in literature during the first part of this research. Since individualisation is a tool for dissuading plagiarism, and the limits of automated marking are more restrictive than the limits of randomisation, the potential exists to make a piece of coursework more reliable through reduced plagiarism without having to convert it to an e-assessment or written examination, both of which would have less potential for high validity. Given the sceptical approach to technological intervention outlined above, this partially-automated approach to assessment is not considered useful simply because it has been proposed; a teaching and learning context is outlined in which it may be useful. The context involves a module in which individual work was used alongside group projects to attempt to take account of individual contribution in the final grade. Since the individual work was necessarily similar to the group project work, the potential for in-group plagiarism was high. The

nature of the topic and the intended learning outcomes did not suit the restraints of a written examination or e-assessment, and so individualised coursework via the partially-automated approach was thought to be potentially very suitable.

The third part (chapters 6-8) describes the development of technology for the proposed partially-automated approach, first by prototyping and later by re-purposing an existing e-assessment system as an assessment-setting system. The implementation of the partially-automated approach in the proposed teaching and learning context is described. An evaluation focuses on whether the specific implementation was successful in achieving its goals, in order to answer the broader question of whether any circumstances exist within higher education mathematics in which the partially-automated approach can be useful.

Essentially, to borrow a mathematical analogy, the second part draws on the findings of the first to propose an 'existence statement' (see Anderson, 1969; p. 29) that a particular novel, apparently unused approach to assessment may exist. The third part then attempts to 'prove' its 'existence' in practice, by finding a particular set of circumstances for the proposed method and examining whether it is more useful in that context than established methods. If so, this project will have provided evidence that there is some merit in including the partially-automated approach in the catalogue of assessment methods available to teachers of mathematics in higher education. This project does not seek to demonstrate that this is the only unrealised novel approach to assessment in higher education mathematics, nor that this is the only way in which a partially-automated assessment approach might be useful (in fact, I would hypothesise that both of these statements are false).

This thesis is in part about the development of my teaching practice (this will be expanded in sections 6.3 and 6.7 in particular), and so is written in first person, although use of first person language is kept to a minimum where it is less relevant to the topic of discussion. My background and motivations, and literature on assessment and e-assessment in higher education mathematics, are introduced in chapter 2. The methodology to be followed at various phases in this project will be introduced in the chapter where the need arises.

1.2 Research questions

1.2.1 Main research questions

1. How is e-assessment used in higher education mathematics, what are its advantages and limitations, and how does it compare to other assessment methods in common use in higher education mathematics in the UK?
2. Can a novel assessment approach be identified which uses e-assessment techniques to provide a combination of potential advantages and limitations that is distinct from that offered by the commonly used methods identified?
3. Is there a particular teaching and learning context in which the proposed novel approach is more appropriate than existing commonly used methods?

1.2.2 Questions related to the particular implementation and its evaluation

The partially-automated approach proposed as a result of research question 2 was trialled in a particular context with a group of students. Six questions are posed in chapter 6 which relate to the evaluation of that implementation and are required to evaluate research question 3.

1. Are the marks particularly sensitive to who is doing the marking?
2. Is the assignment assessing the learning outcomes it was intended to assess?
3. Does the individualised nature of the assignment work to reduce plagiarism?
4. Can the individual work contribute to a view on whether uneven contribution is taking place?
5. Can the individual work contribute to a process for adjusting group marks to take account of individual contribution?
6. What are student views on conventional measures to address uneven contribution?

In addition, the opportunity of surveying students is used to produce data from a different perspective to enable further discussion of research question 1.

7. What are student views on the purpose of assessment?
8. What are student views on automated marking?

1.3 Original contribution

The main contribution of this work is that I propose, develop and trial a partially-automated approach to the assessment of mathematics in higher education, in which e-assessment methods are used to set an assessment which is taken and marked offline. This approach makes it possible to bring automated individualisation to the assessment of higher order skills. In particular, this approach differs from similar approaches taken in statistics (see section 4.5), which use a randomised data sample for individualisation, and from usual approaches to e-assessment in mathematics, where the limitations of computer input and automated marking tend to lead to a focus on mathematical techniques and algebraic equivalence. Given that a review of key projects and reports on assessment in UK higher education mathematics and a systematic review of literature on e-assessment in mathematics did not reveal anything similar, this approach to assessment is claimed as novel in higher education mathematics in the UK.

In addition, this thesis provides a systematic review of literature giving the views of practitioners on the use, advantages and limitations of e-assessment in higher education mathematics. That review did not find any similar published systematic review on this topic, so this is also claimed as an original contribution.

Finally, this thesis attempts to account for a limitation of the literature by providing the results of a survey of the views of those in higher education mathematics who are users of e-assessment but not represented as authors in the literature, and of non-users of e-assessment. The closest found in the systematic review of literature is the survey by Robinson, Hernandez-Martinez and Broughton (2012), which was limited in scope as it focused on lecturers at one university all using the same e-assessment system with first year students (p. 105).

1.4 Overview of chapters

Chapter 2 draws both on reflective evaluation of previous experience and on scholarship to establish relevant background information for this study. A reflection on experiences in higher education mathematics alongside this project details a change in approach to the implementation of learning technology. In order to support

research question 1, a study is made of four key practitioner sources in UK higher education mathematics teaching, learning and assessment, with a particular focus on assessment methods. Finally, a systematic review of literature on e-assessment in higher education mathematics provides practitioner views on the use, advantages and limitations of e-assessment approaches.

Recognising the dominance in the literature of individual enthusiasts, an attempt is made through four interviews to collect the views of users of e-assessment who are not represented in the literature, and of non-users of e-assessment. A questionnaire is used to corroborate or dispute the interview findings among a larger sample. The findings of the interviews and questionnaire are presented in chapter 3.

Chapter 4 discusses the methods of assessment in use in UK higher education mathematics in terms of the five desirable qualities: reliability; validity; fairness and transparency; efficiency; and, usability. This discussion recognises that each method, in general terms, provides access to a different balance of these qualities. This leads to the identification of an opportunity to access a unique balance of qualities through a partially-automated approach, in which e-assessment methods could be used to set an individualised assessment which is taken and marked offline. It is the identification of this opportunity for a partially-automated approach that comprises the key proposal of this thesis.

Chapter 5 outlines a particular teaching and learning context in which the partially-automated approach may be useful. This is a module designed to develop various learning outcomes around graduate skills, and which uses individual work alongside group projects to try to take account of individual contribution in the final marks. The opportunity to use the partially-automated approach proposed in chapter 4 arises during this module. Chapter 5 describes the programme context, module and assessment structure and the topic of the relevant group project, as well as previous experience of running a similar group projects module.

Chapter 6 discusses the design of a framework for evaluation of the implementation of the partially-automated approach in this module based on my development as a teacher of higher education mathematics and on increasing knowledge about the partially-automated assessment approach. This proposes three

strands of evaluation: a second-marker experiment, to examine the reliability and validity of the assessment; a student feedback questionnaire, to examine student views on uneven contribution and plagiarism in the context of this approach, and on assessment and automated marking in general; and, comparison of marks, to examine uneven contribution through individual assignment marks and peer assessment of contribution.

Chapter 7 presents the implementation. First, this details the development of technology to implement the partially-automated approach. The extended timescale of this project means that the underlying technologies changed considerably between early prototyping and eventual deployment. Initial prototyping is described and a system capable of performing the implementation was created, though ultimately this project used a re-purposed existing e-assessment system to set the individualised assessments. Chapter 7 also describes the implementation of the approach proposed in chapter 4 in the teaching and learning context planned in chapter 5, and its evaluation using the method designed in chapter 6.

Chapter 8 gives the results of the evaluation designed in chapter 6 whose implementation was described in chapter 7.

Chapter 9 brings together the findings in the context of the research questions to draw conclusions, discuss limitations and provide further discussion.

2 Background

2.1 Introduction

Moore (2011; p. 14) recommends drawing on a body of knowledge and experience (“theory”), through different contexts, to inform teaching practice. This theory is based on scholarship and also is informed by reflective evaluation of previous experience. Kotecha (2011), similarly, uses a teaching approach which draws on “my past teaching experience; several theories of learning; students' feedback” (p. 5).

It seems appropriate, using this model, to first offer a reflection on the change in approach to implementation of learning technology that has taken place during this project, supported where appropriate by literature, and this is offered in section 2.2. Scholarship on the topic of this thesis takes the form of reading on assessment in general and e-assessment in particular. First, a detailed reading of key sources on assessment in mathematics, statistics and operational research (MSOR) in UK higher education is given in section 2.3. This aims to discuss assessment in the MSOR context and discover the assessment methods in common use. Second, an attempt is made to understand the practicalities, advantages and limitations of e-assessment in the MSOR context. Recognising that advantages and limitations of assessment approaches are not objective truths, an attempt is made in section 2.4 to evaluate these for e-assessment in MSOR via a systematic review of literature, including practitioner opinion and experience.

2.2 Reflection on experience

2.2.1 Introduction

My current research project arose initially from work completed for my Masters dissertation. A change in emphasis and approach has taken place during this project; from naïve enthusiasm for learning technology to a healthy scepticism and more considered approach directed to clear educational need. I don't remember a Damascene moment in relation to this change but a gradual shift based on my activities outside this project.

I began this project in part-time mode in 2004 and alongside this I have had a series of part-time jobs. I started as a computer technician, running websites and databases for statisticians (2004-7). Later I worked for a mathematics professional body, talking to undergraduate mathematics students around the U.K. about career options and what they might do after graduation (2008-10). I worked as a mathematics lecturer in a U.K. university (2008-9), delivering mathematics content for science and business students and computational methods and graduate skills development for mathematicians. I worked to support learning and teaching through technology in a U.K. university mathematics department (2009-10) and I worked on a national project to support curriculum development projects and offer professional development for teaching and learning staff in higher education mathematics (2010-12). Naturally, these experiences have shaped my outlook on pedagogy, and this in turn has influenced this current research.

In this chapter, then, I do not seek to address the main project directly. Rather, I describe some of the other experiences I have had of teaching and of implementing learning technology, and how these have influenced my approach to this research. This essay begins with a critical examination of my Masters dissertation and initial plans for this project. It goes on to describe my experiences of lecturing and the need for deep learning, an appreciation of the role that assessment can play in the development of graduate skills and an understanding of the process of deriving benefit from the implementation of technology. I do not attempt to detail all my previous experience, but to reflect on that which is relevant to the change of

approach that has taken place during my research project.

2.2.2 A solution looking for a problem

My PhD research arose as a continuation of a previous project, my Masters dissertation (Rowlett, 2004a, summarised in Rowlett, 2004b). I completed that dissertation as part of a computing Masters degree following a realisation in 2003 that the technologies used to dynamically generate HTML webpages could dynamically generate MathML code to present mathematics on the web. HTML is the language used to encode content on web pages and uses tags to indicate which parts of a document are headings, paragraphs, links, etc. In a similar way, MathML uses tags to encode mathematical structures. Many HTML webpages are actually output from a script which has drawn together particular page elements (items in your shopping cart, for example) into a unique page for you. Applying this approach to MathML, mathematics can be written which includes dynamically generated elements; in this case to produce questions for an e-assessment system which differ for each user.

I was aware, perhaps from my experience during my undergraduate mathematics degree, that practice is important in learning mathematics. I designed a system to make available to students self-test material on differentiation which would be marked by computer. Using pseudo-randomised constants in the functions to be differentiated, the system was capable of generating something in the region of nine million distinct questions from a dozen or so basic question types. A student using this system could reasonably expect to be able to practice for as long as they liked without running out of new questions to attempt.

The dissertation itself (Rowlett, 2004a) has a strong focus on technology with chapters explaining MathML and dynamic webpage technologies and the main focus of the work being on applying the latter to the former to produce a computer-aided assessment system. A chapter on 'Mathematics and CAA' (p. 22-25) does mention some aspects of how people learn but has as its main focus issues arising from using technology to present mathematics assessments, including problems inputting mathematical notation and limitations of automated marking.

A user evaluation questionnaire was attempted but this gathered few responses. The evaluation chapter (pp. 45-49) focuses instead on whether the system was generating sensible mathematics without error, based on limited use by students. Although users were considered in the concluding chapter (p. 50), for example there is some concern expressed about “the small scope of the student usage evaluation”, the main conclusion is that “utilising the power of MathML through its dynamic manipulation has some merit when applied to the production of pseudo-randomised mathematics questions”. This seems to mean only that it is indeed possible to use an e-assessment system to set valid questions using MathML and correctly mark them.

Exhibited in this work is clear pride in the efficiency of the approach (for the assessor); the system was capable of generating a relatively large quantity of 'output' for the amount of 'input' required. The starting point of that research was noticing what could be done with the technology and implementing this, simply because it was possible, without reference to wider pedagogic theory. In a way, this was quite reasonable; I was applying what I had learned during my computing degree to the interesting area of computer-aided assessment. In that sense I remain proud of the work; but at the same time it now seems hopelessly naïve, lacking any in depth focus on the educational aspects of what was being attempted.

I began my PhD with a similar approach. I was encouraged early on in this (2004/5) to write an account of my plans (Rowlett, 2005) and in this I describe my research as continuing the work started during my Masters dissertation and “to investigate the application of other [similar] technologies to mathematics teaching and learning” (p. 35). Again, the focus seems to be on what can be done with technology first and what might be needed second.

2.2.3 Deep learning of mathematical concepts

In 2008 I began lecturing a basic course in mathematics for business and science students who did not hold a recent qualification which included mathematical topics relevant to their degree. Instinctively, I wanted to encourage the students into some understanding of the concepts behind the mathematical syllabus, rather than just treating mathematics as a set of techniques to be memorised. As a mathematician, I

have a poor appreciation of how one might learn (certainly, remember) a technique without first understanding it. However, the students I was teaching were to use mathematics as a tool, a means to an end. In their home disciplines they would have problems which need to be solved using mathematics and they were looking to be taught how to apply relevant techniques. In circumstances such as these, I wondered, is it sufficient to teach the procedure of applying the technique?

In the context of my e-assessment project, a system which can generate millions of instances of a limited range of questions might be seen as supporting the latter type of learning. It helps students to memorise and practice the technique without necessarily supporting their understanding. But is that a problem?

Entwistle and Ramsden (1983) contrast two main types of learning: “deep” (or “meaning”) and “surface” (or “reproducing”) (p. 193). The deep approach being “internal”, focused on the content of the article or problem and the knowledge, experience and interests of the learner. The surface approach is “external”, focused on the task and its requirements, with material “impressed on the memory for a limited period” and “no expectation that the content will become a continuing part of the learner's cognitive structure” (p. 195).

Raine (2005) regards the way mathematics has been taught to science students as procedural, “by constant repetition and coverage of all possible variants of a given problem, eventually to the extent that mathematics becomes indistinguishable from pattern matching” (p. 14). Fuson, Kalchman and Bransford (2005) suggest this procedural way of teaching mathematics “often overrides students' reasoning processes, replacing them with a set of rules and procedures” (pp. 217-218). These descriptions certainly sound like they would fit a massively randomised e-assessment system. Since the computer is applying a set of rules and procedures for pattern matching, we can hardly expect that repeated use of the system would encourage anything deeper from students. So does this sort of learning suit the requirements some students have to just learn to apply mathematical techniques?

Fuson, Kalchman and Bransford argue that focusing the instruction on procedural knowledge is ineffective and causes a disconnect from the meaning behind the mathematics, evidenced by students failing to correct erroneous answers which are

clearly unrealistic. Raine remarks that the surface approach works “until you change the context and start asking for applications of the techniques in unfamiliar surroundings” (p. 14). A study in which students could reproduce a desired technique well but could not adapt this knowledge to unfamiliar circumstances is reported by Dreyfus (2002). There is also some suggestion that deep understanding can help both in the retention of what is learned and confidence in the subject (Entwistle and Ramsden). Fuson, Kalchman and Bransford even attribute some students' dislike of mathematics as a subject to focusing the instruction merely on procedural knowledge.

Of course, the purpose of science and business students learning mathematics is that they can retain this knowledge and apply the techniques to unfamiliar scenarios in their home discipline and wider career. If such students are dissuaded from applying mathematics when necessary because they have forgotten the techniques, lack confidence to apply them or even because of dislike of the subject then this is extremely problematic. Looking at it this way, then, the surface approach seems to be especially poor preparation.

My reading around this subject led me to be concerned that I was simply training my students in the mathematical techniques so they would acquire, as Dreyfus puts it, “the capability to perform, albeit much slower, the kind of operation which a computer can perform” (p. 28). Beyond the immediate experience, this investigation led me to believe that the approach explored during my Masters research was similarly flawed. If encouraging students to practice large numbers of out-of-context, little-varying, self-test questions is implicitly encouraging procedural, and therefore surface, learning, then this approach will discourage students' ability to retain and apply their mathematical knowledge.

2.2.4 Development of graduate skills

In 2009 I was to give a module for second year mathematics students which aimed to use mathematical project work to develop graduate skills, including time management, working in small teams and communicating using reports and presentations. These aims, including a syllabus which did not intend to develop any

particular mathematical topic, would require a very different approach to assessment.

Beevers and Paterson (2002) describe “key skills” as “what is left after the facts have been forgotten” (p. 51). Challis et al. (2002) define a subset of key skills as “transferable” (p. 80) and say that as well as academic knowledge, professional mathematicians require these skills to “use their knowledge effectively” (p. 89). They say it is “incumbent on us, as teachers, to help our students to learn and develop these skills” (p. 80). Lowndes and Berry (2003) agree, saying that employers have “voiced their perceptions/criticisms that students/graduates are technically competent but lack professional skills, awareness of business issues, communication skills [and] problem solving skills” (p. 20). There is clearly a need to develop these skills, but how is this achieved?

Hibberd (2005) notes that “much of the teaching and learning in an undergraduate mathematics curriculum is provided by traditional lectures and problem workshops and assessment is dominated by examination” (p. 5). This view of assessment of undergraduate mathematics in the U.K. is well supported by the findings of Iannone and Simpson (2012a) (more in section 2.3). MacBean, Graham and Sangwin (2001) note that “many people” view mathematics as a subject in which “something is either right or wrong” so that it is “difficult to discuss or debate and ... not open to differing opinions” (pp. 1-2). This sort of teaching and assessment, and this view of mathematics, is well suited to short problem questions with well-defined correct answers of the sort that might be marked by computer.

Hibberd (2005) suggests that these teaching and assessment methods are “strong” for “the attainment of knowledge” but make “more limited contributions to other elements” (p. 6). Thomlinson, Robinson and Challis (2010a) recognise that e-assessment, with its “rapid feedback”, can “promote engagement” in the first year of a degree, but that the use of closed questions required for rapid computer marking limits the potential to develop graduate skills (p. 126). If widely used methods limit graduate skills development, is there a kind of assessment which is more suitable?

Waldock (2011) argues that graduate skills can be developed by using alternative methods of assessment which encourage skills development alongside mathematical content, in a way that traditional assessment methods do not. Reporting on interviews

with senior staff in U.K. university mathematics departments, Iannone and Simpson (2012a) say that “some justified alternative forms of assessment as more realistic of the kinds of tasks students would encounter in later employment” (p. 13). Specifically, Hibberd (2002) recommends that group project work can “[lead] students into a more active learning of mathematics, and an appreciation and acquisition of associated key skills” (p. 159). A case study of Iannone and Simpson (2012a) supports this, reporting the move away from examinations towards more project work at one U.K. university mathematics department as part of a drive to offer “the development and assessment of a wider range of skills” (p. 8).

Traditional methods of assessment, using short problems with well-defined solutions, are well suited to use of e-assessment and computerised marking but less well suited to the development of graduate skills. Challis et al. advocate that the development of transferable skills “should be embedded in the mathematics curriculum” (p. 80) otherwise, they say, skills-based assignments risk being seen as “an ‘add on’” rather than “an integral part of a mathematician's life” (p. 90). It seems that the approach taken in my Masters research, involving automated marking, is particularly at odds with the objective of embedding graduate skills development throughout the degree programme.

2.2.5 The pattern of the technology enthusiast

In 2009 I began working to support the mathematics curriculum through technology in a U.K. university mathematics department. Having been concerned that my e-assessment system may encourage surface learning and preclude the type of assessment that most favoured the development of graduate skills, my experiences in this role most of all made me question my 'wide-eyed' approach to using technology.

Often, technology is introduced out of general enthusiasm. I certainly saw that in myself and others during this role, and it lay at the heart of my Masters dissertation project. Apart from my own experience, I have met a number of technology enthusiasts and heard several seminars on projects that seem to exhibit the same approach: it looks interesting, let's try it and see what it can do; we can worry about what we are trying to achieve later.

I became involved with a project to conduct whole class quizzes through the use of response system technology (sometimes called electronic voting system or 'clickers'). Audience members enter individual answers to questions via a remote device and these responses are displayed collectively to the group. This technology is thought to bring two-way communication in large lectures and provide a beneficial active learning opportunity to every member of the audience (Simpson and Oliver, 2007). We used the technology in a large introductory applied mathematics module for regular quizzes to encourage students to keep up-to-date with lecture content and not simply save it all up to revise at the end. Each quiz was conducted, answers were given in class and students were encouraged to view worked solutions on the web later.

The lecturer felt that the technology was useful and he had received informal positive feedback from students. This encouraged him to plan to repeat the process the following year. An evaluation of students' reported use of the feedback they received during quizzes, however, suggested that the technology was only benefiting those students who were more likely to engage in any case (Barton and Rowlett, 2011). Essentially, students who reported that they were encouraged by the quizzes to review module materials said that they kept up-to-date equally well with other modules where clickers were not used. Students who were not encouraged to engage were struggling to keep up with this and other modules equally.

Studies which reported a positive benefit for students were those that used clickers to drive an active change in teaching practice (Dufresne et al., 1996; Crouch and Mazur, 2001), while those reporting no evidence of benefit were controlling for other factors, such as teaching and learning method, when introducing the new technology (Kennedy and Cutts, 2005; King and Robinson, 2009). The technology introduction in our experiment was not accompanied by an educational change, such as Crouch and Mazur's peer instruction driving a more active style of learning, and so the findings of the evaluation should perhaps not have been a surprise.

This technology was applied more or less in isolation, without driving a change in educational approach. It took time in class and effort on behalf of the lecturer and students, with positive effect on engagement with module materials, the stated goal, only reported by students who self-reported as being more likely to engage anyway.

Despite the positivity of the lecturer and the students he spoke to, an uncritical, unchanged repetition of use of the technology could be unhelpful or even detrimental. Use of the technology did provide the lecturer with valuable data about what the students did and did not understand in the form of their answers. If that is an acceptable goal, and the downside of time taken in class an acceptable price to pay, then the technology introduction may have been a success. As the desire was in fact to improve student engagement, and ultimately student learning, questions ought to be asked about implementation before the technology is used again. Learning to connect plans for technology implementation to clear educational goals and later evaluate against those goals is an important step which was missing from my Masters dissertation.

2.2.6 Discussion

At the start of this process I was a naïve enthusiast willing to apply technology simply to see what can be done. Through an investigation of teaching methods to develop deep learning I came to worry that using e-assessment to generate large numbers of out-of-context, little-varying, self-test questions might give students the wrong message about the aims of assessment and encourage surface learning. This mode of learning can be detrimental to students because without understanding, they may not be able to apply techniques to unfamiliar circumstances and might struggle to retain what they have learned. Through designing a group project task to develop graduate skills I came to understand that producing graduates equipped for the challenges of life requires more complex tasks, and not just problems which are unambiguously right or wrong, to be embedded throughout the curriculum, limiting the effective range of e-assessment using automated marking across a degree programme.

Through investigations into learning technology I came to understand the pattern of the technology enthusiast and recognise this in myself. This is where the naïvety of my earlier approach is laid bare. A new technology is introduced, perhaps without a particular aim and fuelled by individual enthusiasm. This then receives positive feedback from those students who used it, perhaps masking disengagement by a silent minority who are disadvantaged, which in turn fuels further uptake of the

technology.

Actually, to expect an automatic improvement of student learning simply by replicating a current approach using technology is not encouraged by my experiences. Technology introduction ought to be put in a context of educational need and implemented accordingly. Educational technology may not produce a benefit simply by its introduction, but a benefit may derive from a change of approach driven by use of the technology. Approaching a curriculum development without considering potential educational need seems to me, now, to be a thoughtless approach. As such, a technology intervention should follow one of two patterns: to drive a change of practice which is beneficial to student learning, or to replicate a current approach more efficiently without detrimental effect and with no expectation of effect on student learning, perhaps in order to release staff time for another activity to positively influence student learning.

The approach taken when producing an e-assessment system for my Masters dissertation suffered from these problems: it was implemented with enthusiasm simply because it could be done and not to address any educational need; it encouraged surface learning; it precluded the sort of assessment that assists with graduate skills development; and, just because I was positive (and, given more time I may have found some students who were positive also) does not mean that the technology was benefiting anyone involved.

My focus in the years since I started my PhD project has shifted considerably away from enthusiastic implementation of e-assessment technology to one looking at the effect of using such technology and when and where it can be implemented to address an established need and produce an educational benefit. In general, my approach to using technology in education has become much more sceptical. I listen to people who say they are implementing some technology, or asking me whether I am planning to introduce something, and my first question is: what is the educational need and how could technology address this better than the current approach? Otherwise, I won't waste everyone's time.

2.3 Assessment in higher education mathematics

2.3.1 Introduction

Assessment in mathematical sciences and the methods used are examined through a detailed study of selected key sources. The relevant QAA Benchmark Statement¹ (QAA, 2007) recognises that assessments in mathematics “differ substantially” from those in other subjects (p. 20). For this reason, this section draws on sources on assessment specifically in higher education mathematics.

Challis, Houston and Stirling prepared sets of resources aimed at university staff (2004a) and students (2004b), published in 2004 by the Learning and Teaching Support Network Subject Centre. These were based on a set of generic Learning and Teaching Support Network resources but designed “to interpret and add to them in such a way as to bring out their relevance and usefulness in the context of Mathematics, Statistics and Operational Research (MSOR)” (2004a, p. 2).

Iannone and Simpson (2012b) ran a project, Mapping University Mathematics Assessment Practices (MU-MAP), supported by the National HE STEM Programme in 2011-12, which aimed “to examine the current state of assessment in our undergraduate degrees”. They designed this “not only to give a broad overview of practice, by looking across our higher education institutions, but also to have an eye to the future and alternatives” (p. iii).

As well as these resources with a specific focus on assessment, two recent sources devoted to higher education mathematics were consulted for their sections on assessment.

Robinson, Challis and Thomlinson (2010), working as part of the More Maths Grads project in 2007-10, aimed to “explore and review the way mathematics courses are working” in four institutions through questionnaires, group and individual

¹ The QAA Benchmark Statement for Mathematics, Statistics and Operational Research (MSOR) (QAA, 2007) aims to “provide a means for the academic community to describe the nature and characteristics of programmes” and “represent general expectations about standards for the award of qualifications at a given level in terms of the attributes and capabilities that those possessing qualifications should have demonstrated” (p. ii).

semi-structured interviews with students and staff. This was “partly concerned with curriculum content” and also with (Challis, Robinson and Thomlinson, 2010; p. 5)

understanding and improving the kind of experience we provide for our students of mathematics, how we teach them, engage them and support them and recognise their aspirations, how they feel about that experience, and what the implications are of all that.

Particularly the chapters by Thomlinson, Robinson and Challis (2010a, 2010b) focus on assessment.

Cox (2011) offers a guide to practice for new lecturers in “mathematics and related subjects”. Cox hopes to avoid being “idiosyncratic, representing the single view of the author about teaching matters” by basing his book on “the production of materials for the highly successful Maths, Stats & OR (MSOR) Network Induction Course for New Lecturers” and thus “taking advantage of the wide range of input available to the Network”. Cox aims “to provide support for any new MSOR lecturer” and targets “the best interests of the students while being practical about the pressures on their teachers” (p. xi).

This seems to be a suitable mix of approaches to provide useful coverage of the topic. Challis, Houston and Stirling (2004a, 2004b) provide a link to the general assessment literature, filtered for relevance to MSOR, Iannone and Simpson (2012a, 2012b, 2012c) review the current state of mainstream and alternative assessment practice, Robinson, Challis and Thomlinson speak to staff and students about their experiences and Cox provides guidance to train staff.

This section is split into two parts: the first is about assessment in general; the second discusses specific methods of assessment.

2.3.2 About assessment and its evaluation

2.3.2.1 What is assessment?

Cox offers the following definition of assessment (p. 149):

The measurement of the extent to which students have met the learning objectives of a course of study.

Cox says that “assessment is a major part of individual teaching and of departmental administrative affairs” (p. 145) and “in practice most teaching and learning is

assessment driven” (p. 149).

2.3.2.2 The purpose of a mathematics degree

The QAA Benchmark Statement expects certain subject-specific skills from all graduates, and expects “most of these” to be “formally assessed at some stage” (QAA, 2007; p. 15). The abilities listed are:

- “to demonstrate knowledge of key mathematical concepts and topics, both explicitly and by applying them to the solution of problems”;
- “to comprehend problems, abstract the essentials of problems and formulate them mathematically and in symbolic form, so as to facilitate their analysis and solution, and grasp how mathematical processes may be applied to them, including where appropriate an understanding that this might give only a partial solution”;
- “to select and apply appropriate mathematical processes”;
- “to construct and develop logical mathematical arguments with clear identification of assumptions and conclusions”;
- “where appropriate ... to use computational and more general information technology (IT) facilities as an aid to mathematical processes and for acquiring any further information that is needed and is available”;
- “to present their mathematical arguments and the conclusions from them with accuracy and clarity” (p. 15).

The Statement says that particular branches of MSOR will have “other subject-specific skills that are relevant to those particular branches” (p. 15). This also says graduates will have “highly developed skills of a more general kind” (p. 13) and that these will be “honed by their experiences of studying MSOR subjects” (p. 16). The Statement stresses “a general ethos of numeracy and of analytical approaches to problem solving”, and the importance of “the general skill of transferring expertise from one context to another” (p. 16).

The Statement suggests that graduates will possess the following general skills (p. 16):

- “general study skills”;

- able “to learn independently”;
- able “to work independently with patience and persistence, pursuing the solution of a problem to its conclusion”;
- “time management”;
- “organisation”;
- “be adaptable, in particular displaying readiness to address new problems from new areas”;
- “able to transfer knowledge from one context to another”;
- able “to assess problems logically and to approach them analytically”;
- “highly developed skills of numeracy”;
- “general IT skills, such as word processing, the ability to use the internet and the ability to obtain information, always exercising these skills in a responsible way and taking care that sources are referred to appropriately”;
- “general communication skills, typically including the ability to work in teams, to contribute to discussions, to write coherently and to communicate results clearly”.

Where appropriate, this says, graduates will “have knowledge of ethical issues”. The Statement recognises that these skills “enhance the general employability” of graduates (p. 16).

The QAA Benchmark Statement also says that assessment “is not necessarily restricted to the assessment of mathematical knowledge and understanding alone” and that programmes will also assess, for example, “the ability to use mathematical ideas in the context of an application” or “the ability to communicate effectively” (QAA, 2007; p. 18).

2.3.2.3 Purposes of assessment

Challis, Houston and Stirling (2004a) give six “purposes of assessment” (p. 44):

1. “to give a licence to proceed to the next stage or to graduation”;
2. “to classify the performance of students in rank order”;
3. “to improve student learning”;
4. “to inform teachers of the strengths and weaknesses of the learners and of

- themselves so that appropriate teaching interventions may be employed”;
5. “to inform other stakeholders – society, funders, graduate schools, potential employers”;
 6. “to encourage learners to take a critical-reflective approach to everything that they do, that is, to self-assess before submitting”.

Cox lists the following as possible purposes of an assessment (pp. 149-150), based in part on Brown (2001; p. 10):

- “to judge the extent to which knowledge and skills have been mastered”;
- “to monitor improvements over time”;
- “to diagnose students' difficulties”;
- “to evaluate teaching methods”;
- “to evaluate the effectiveness of the course”;
- “to motivate students to study”;
- “to predict future behaviour and performance”;
- “to qualify students to progress”.

Cox says that whatever its purposes, assessment is “primarily a measure of student learning”. He warns that these purposes must be used “carefully” in balance. For example, a coursework assessment used to encourage student learning may not be very effective at predicting students' future performance “if only because there is no certainty that it is solely the student's work” (p. 150).

Note that the first list contains “to improve student learning” and the second “to motivate students to study”. Cox says that one of the important points of assessment is that “students best learn mathematics if they are actively engaged in the process of doing mathematics”, and that this is best achieved in an examination (p. 149). Challis, Houston and Stirling (2004a) agree that active learning as a means of promoting a deep approach should be promoted, but do not so restrict the range of assessment methods (p. 44).

2.3.2.4 Assessment of graduate skills

Challis, Houston and Stirling (2004a) report “the wide acceptance of the [QAA] benchmark statement” in MSOR and therefore “some general acceptance of the value

of key skills²” (listed in section 2.3.2.2), while pointing out there is “less agreement over how to embed or to assess students' development of them”. They also point out that employers have (p. 15)

consistently said they value transferable or key skills, and it behoves us as course designers to bear in mind the future needs of mathematics graduates working in industry, whether as a mathematician or in a more general capacity.

In section 2.2.4, the issue of graduate skills development in teaching was discussed, with the idea that this must be embedded throughout the curriculum. Challis, Houston and Stirling (2004a) also report this view and, having acknowledged that “many students now work strategically, responding only to the pressure of assessment” (p. 7), they say (p. 16)

others may feel that integrating the skills assessment distorts the integrity of their award. However, working on the principle that if you want to convey to students that you value something then you have to assess it, this issue must be addressed.

They suggest communication skills can be developed through “communicating the answer to a range of people and not just peers, in suitable form(s) (e.g. report, article, poster, oral presentations, etc.)”. They say “the mathematical modelling process provides an ideal vehicle for developing and assessing the full range of communication skills, from understanding the problem, to communicating validated conclusions, perhaps to non-experts” (p. 16). They also suggest mathematical modelling “covers the whole territory” of problem-solving, providing “an ideal vehicle both for skills development and assessment” (p. 17).

Challis, Houston and Stirling (2004a) acknowledge that group work only receives “one very brief mention” in the QAA Benchmark Statement but say that (p. 17)

in the world of work most problem solving involves teamwork, and while a mathematician will bring special skills to a team, she or he will also have to function as a member of that team and to be aware of the issues involved in that.

Again, they cite modelling as “an ideal vehicle for developing group working skills” (p. 17).

² “key skills”, “transferable skills” and “graduate skills” are used here interchangeably.

2.3.2.5 Learning outcomes and alignment

Challis, Houston and Stirling (2004a) define learning outcomes as “statements of the things it is intended that students will be able to demonstrate or do when they complete the programme or a constituent module” (p. 43). They say that the “most useful piece of advice” on assessment for lecturers is “first of all write the *learning outcomes* of a module *very carefully*”, since “thoughtful learning outcomes are the key to good assessment” (p. 42; emphasis original). This is because, for a clear set of outcomes representing a concise statement of what the students will learn, a module is made up of teaching, learning and assessment methods which are designed to “match the chosen learning outcomes” (Kahn, 2002; p. 93).

2.3.2.6 Formative and summative assessment

Cox has formative assessment as “designed for developmental purposes” and not contributing to students' marks or grades, “allowing students to make mistakes without penalty” (p. 150). He warns that it is possible “students won't do such work because it does not attract marks”, meaning that it is possible that only the more motivated students will complete formative work (p. 150). Cox says summative assessment is “designed to establish students' achievement at stages throughout a programme and normally contributes to their marks and grades” (p. 150). Challis, Houston and Stirling (2004b) agree that summative assessment “attempts to measure the extent to which a student has achieved the learning outcomes of the programme or module” (p. 4) and that it is used to produce “a final mark at the end of the process” (p. 7). Challis, Houston and Stirling (2004a) say that “quite often” the same assessment tasks are used for formative and summative assessment (p. 5).

2.3.2.7 Marking criteria

An assessment is accompanied with a set of marking criteria or a mark scheme. Cox says “the essentials” of good marking criteria are that they:

- “match the assessment tasks and learning objectives”;
- “enable consistency of marking”;
- “can pinpoint areas of disagreement between assessors”;

- “can be used to provide useful feedback to students” (p. 166).

On consistency between multiple markers, Thomlinson, Robinson and Challis (2010a) say defining clear instructions for “different markers of varying skills”, in order to improve reliability, “constrains the type of questions being asked since students need to be steered to produce answers in a reasonably standard form” (p. 123).

2.3.2.8 Plagiarism

Challis, Houston and Stirling (2004b) define plagiarism as “misrepresenting someone else's work as your own” (p. 17). Challis, Houston and Stirling (2004a) highlight a variety in definitions of plagiarism, with some being “more explicit about the dishonesty” and others putting “greater emphasis on the acknowledgement of sources”. The latter is a more subtle issue, for example including the extent to which information given in the lectures - “the ‘common knowledge’ of the subject” - must be acknowledged, or whether students properly acknowledge when they are quoting “large amounts of text” from other sources. There are considerable issues here, particularly around communication of good practice and expectations to students (p. 23). Cox warns that plagiarism can distort the view of the lecturer on whole-class progress (p. 183).

Iannone and Simpson (2012a) note that plagiarism can be hard to identify in short problem questions, as “it would be quite natural for students to use similar methods and even similar variable names in their solutions to the same problem” (p. 12). Cox describes the situation thus (p. 182):

Particularly in techniques-based topics two completely independent solutions to the same problem can look virtually identical. The only real give away is when a number of scripts come in together with identical silly mistakes, and even then, unless someone coughs it is difficult to assign responsibilities.

Challis, Houston and Stirling (2004a) also discuss collusion. They say that we “ought to commend” the practice of students discussing “a piece of work they all have to do”. They say that students “might think through the issues together and come to some joint conclusions”. They say that “in group work this is explicit and uncontroversial, as the output will be acknowledged as coming from the group”

(although, re. “uncontroversial”, see the discussion under section 2.3.3 about assigning individual marks for group work), but that in cases where the work is individual “there is frequently a grey area here”. They say few would object if students “all contributed to the discussion and the individual write ups are different”. They say that in “essay-type material” (p. 24)

it is reasonably clear that independent write-ups will be different, but for analysing data or solving a mathematical problem this is less clear: the theory will constrain independent correct work to be very similar, at least for students of moderate ability.

Cox says the issue of plagiarism is “a hot topic nowadays” but “somewhat overcooked” (p. 182). Challis, Houston and Stirling (2004a) say that for a student to plagiarise “is not sensible” because (p. 23)

if there is a purpose in asking the student to do the work it will be to promote or assess the student's learning, and copying contributes nothing to that purpose.

However, Iannone and Simpson (2012a) report concern about a “thirst for grades” driving “plagiarism or collusion with coursework” (p. 13).

2.3.2.9 Evaluation of assessment

Assessment can be evaluated in terms of the following qualities: reliability, validity, fairness (QAA, 2007; p. 18), efficiency and usability (Challis, Houston and Stirling, 2004a; p. 6). Definitions of these terms are provided below.

Reliability

Challis, Houston and Stirling (2004a) say “reliable assessment measures are objective, accurate, repeatable and correctly marked and recorded” showing “no bias between examiners” (p. 5). Cox defines reliability as meaning that “the outcome of the assessment is consistent for students with the same ability, whenever the assessment is used, whoever is being assessed, and whoever conducts the assessment” (p. 151).

Challis, Houston and Stirling (2004a) say that in “traditional mathematical assessment there is a feeling that examiners are highly likely to agree, since a piece of mathematics is either right or wrong”. They say that “reliability has been an issue for more discursive disciplines to worry about but not us” (p. 19). However, complete

reliability may not be possible. Cox warns about multiple markers, saying that “any scheme that is sufficiently detailed to always guarantee reliability is almost certain to be complex and unwieldy, or to render the question so anodyne as to pose little challenge” (p. 151). Challis, Houston and Stirling (2004a) say (p. 19),

now that many MSOR courses do include some learning outcomes with a more discursive flavour, perhaps encompassing project, modelling and related work, and the requirement for instance for report writing, then we must face this issue.

Cox also says assessment must be objective. He says “nowhere is it more important to balance professional position and ordinary human regard for students”; also that assessors must “put aside any personal feelings to ensure objective decisions on the outcomes of assessment” (p. 147). Challis, Houston and Stirling (2004b) say anonymity is used to avoid examiner bias “towards or against a particular candidate”. They say this is used “for the most important assessments (e.g. exams)”, but acknowledge that in some circumstances “this anonymity is impractical” (p. 5).

Validity

Challis, Houston and Stirling (2004b) define validity by saying simply: “assessment should test what it is supposed to test” (p. 5). Cox defines validity as “the requirement that the assessment measures attainment of the learning objectives set” (p. 151). Challis, Houston and Stirling (2004a) say that valid assessment “must assess all the things we expect students to learn which are considered to be of value” (p. 6), and that “assessment should test whether a student has achieved the goals of the module, course or programme” (p. 19).

Challis, Houston and Stirling (2004a) say “there has to be alignment between learning outcomes, methods of learning, assessment methods and assessment tasks” (p. 6). They say that “while the three-hour unseen written examination is commonplace, there are considerable limitations to what can be achieved and measured in such a format; indeed the same is true of traditional 'right or wrong' coursework”. They point out that issues of validity are particularly relevant when assessing “mathematical modelling activity” (p. 19).

Fairness and transparency

Cox says assessment must “be 'fair' to students and give them every opportunity to show... what they can do” (p. 145). Cox says questions must be “carefully explained and unambiguous” (p. 148), and be “interesting” without being “a 'trick question'” (p. 149).

Related is the issue of transparency. Challis, Houston and Stirling (2004b) say students “should be told how [they] will be judged” (p. 5). They say this is not detail of questions but relates to: how much the assessment contributes to the module and to the overall degree assessment; what students have to do to pass; arrangements for resits, special circumstances, etc. (p. 6). Cox also highlights the need for “clarity and precision about what is expected of the students and how this will be measured”, but balances this against over-specifying, which would be “tantamount to telling the students precisely what the assessment will be” (p. 147).

Efficiency

Challis, Houston and Stirling (2004a) say that “with enough resources reliability can be improved” but this may not be possible “due to resource constraints” (p. 6). Cox says assessment must be “practical”, explaining that “all assessment is a compromise between ensuring the validity and reliability of the assessment and working within limited resources” (p. 147).

Cox also highlights the need for providing a reasonable workload for students. This means providing “not too much (or too little) coursework” and “questions and examination papers need to be of a reasonable length and difficulty” (p. 148).

Usability

Challis, Houston and Stirling (2004a) define usability as “the *usefulness* of assessment to each of the internal stakeholders – students, lecturers, managers, and each of the external stakeholders – postgraduate schools and employers” (p. 6; emphasis original). Specifically, each of these are said to be interested in:

- students: feedback on learning and marks for progression and award;
- lecturers: information about the progress of students;
- managers: are “keen to observe good quality”;

- postgraduate schools: details of a student's whole profile;
- employers: “assurance that the student knows the subject, perhaps even specialised topics, and whether the student has good skills” (p. 6).

Evaluation of an assessment

The QAA Benchmark Statement says assessment criteria will refer to “validity, reliability and fairness, assessing appropriate aspects of knowledge and skill, supporting the aims of learning and teaching, providing opportunities for feedback whenever appropriate, and being open to external scrutiny (for example, by external examiners) as appropriate” (QAA, 2007; p. 18). Challis, Houston and Stirling (2004a) say that (p. 42)

it should be the aim of every examiner to set examinations and assessment tasks, and to devise assessment schemes that are as reliable and as valid as possible given the constraints of affordability, and to be of use to every stakeholder in the whole process of educating students.

Cox reports that evaluation of assessment is completed by “the external examiner”, that “most modules will usually have a moderator to check the examiner's work” and that there is often departmental “scrutiny of the outcomes of the assessment” (p. 146). Challis, Houston and Stirling (2004b) say that the external examiner will “oversee the fairness and quality of the whole process” (p. 7). Cox says “at a superficial level, you need to ensure that your results are 'reasonable'. Your examination board Chair or Examinations Officer will soon let you know if they are not!” (p. 186).

2.3.3 Methods of assessment used

2.3.3.1 Introduction

The QAA Benchmark Statement says that (QAA, 2007; p. 18)

learners are likely to meet a range of methods of assessment, depending on the aims and learning outcomes of the programme or module and on the knowledge and skill being assessed. Assessment will be carried out according to context and purpose.

Challis, Houston and Stirling (2004a) say that “there are appropriate learning methods, appropriate assessment methods and appropriate assessment tasks for

different learning outcomes, and it is important to align these in a sensible way”.

They say that proper thought given to learning outcomes “requires the introduction of different methods of assessment, methods that are better than written examinations at assessing some learning outcomes like, 'On successful completion of this module students should be able to demonstrate effective teamwork skills'” (p. 42).

Several of the sources consulted discuss specific methods of assessment. In particular, Challis, Houston and Stirling (2004a, 2004b) give guidance on different methods, Iannone and Simpson conducted an extensive survey of current assessment practice in mathematics departments in England and Wales (2012a), and collected a series of anonymised case studies based on current practice through an opportunistic sampling method (2012b). Focus is kept on non-e-assessment approaches, given the more systematic treatment of literature on e-assessment methods in section 2.4, though actually e-assessment approaches were not widely discussed in the sources consulted here.

Challis, Houston and Stirling (2004b) list eight assessment types (unseen timed written examination, open book exam, seen timed written examinations, coursework, multiple choice exams, oral exams, presentations and projects or dissertations) (pp. 8-11) and say these are “only a selection” which includes “the most common ones in MSOR” (p. 12). Also included in this section are posters (Iannone and Simpson, 2012b; pp. 39-40) and peer assessment (Challis, Houston and Stirling, 2004a; p. 27).

Challis, Houston and Stirling (2004a) also mention self-assessment and say that this is “perhaps one of the most important skills that higher education can impart to its students” (p. 27), although it is not specifically highlighted in the other sources consulted. They say such reflective tasks in MSOR would “normally” be “for formative purposes” to “promote deeper learning” rather than “to rank students” (p. 27). This is noted as important, then, but not included in the list of assessment methods below.

2.3.3.2 Written examination

Of 1843 modules examined by Iannone and Simpson (2012a) at UK universities, “over one quarter” (535) were “assessed entirely by closed book examination” and “nearly 70%” (1267) “use closed book examinations for at least three quarters of the final

mark” (pp. 4-5). Iannone and Simpson (2012a) suggest that most departments are content with the dominance of unseen, closed book written examinations, but that alternatives have been introduced in some cases, particularly to introduce tasks “more realistic of the kinds of tasks students would encounter in later employment” (p. 13).

Cox says that the unseen written examination is usually “closed book”, save perhaps for “formulae sheets” (p. 159). Iannone and Simpson (2012a) report use of “open book” examinations, in which “anything from statistical tables to textbooks and other materials” may be taken in to the examination (p. 10). Challis, Houston and Stirling (2004a) also outline “seen” examinations, in which students are “given the questions some time in advance” (p. 45). Challis, Houston and Stirling (2004b) describe multiple-choice examinations, “paper- (or maybe computer-based) exams which ask you to pick the right (or best) option from typically four options” (p. 10).

Challis, Houston and Stirling (2004a) say that “written examinations are pretty *reliable*, but they only have limited *validity*” (p. 6; emphasis original). They say that (p. 6)

the more complex the task, the less likely it will be that a written examination will be a valid means of assessment; the more we simplify the assessment, the more reliable it becomes, but at the expense of validity.

Challis, Houston and Stirling (2004b) say that problems-based exam questions are “in many ways the least complex form of assessment - and one of the most ‘valid’ (and reliable to mark)” (p. 9). Challis, Houston and Stirling (2004a) say that this method is “very reliable in assessing particularly the lower order cognitive skills, but not useful for assessing extended investigations” (p. 45).

Cox, however, says that “a properly set unseen examination of sufficient duration can assess **any** cognitive skill” (p. 158; emphasis original). He says that open-ended questions are possible in an examination in some contexts via “unseen generalizations and abstractions”, though he says that such questions should be “less demanding” in examination questions than in coursework (p. 148), which seems to limit the validity of this assessment method. Similarly, he says that deep learning is “much easier to assess in projects or coursework” but that an examination question might ask

students “to describe how they would approach a particular previously unseen problem, without actually solving it” (p. 148).

Cox points out that “examination conditions” can “usually guarantee” that the assessment provides “reliable results”, but “take away coursework cannot” (p. 151). In particular, Cox says that “supervised examinations” provide “little possibility of cheating” (p. 182).

2.3.3.3 Coursework

Two types of coursework are reported: frequent short coursework, much like a take away exam paper; and, longer, in depth pieces of work. The latter might be the result of project work.

Cox says that “unsupervised coursework” is “problematic” in relation to plagiarism (p. 182). In their interviews, Iannone and Simpson (2012a) note that “often people explicitly raised concerns about weekly homework sheets and how they could be sure that the work represented the students' own efforts” (p. 12). Iannone and Simpson (2012b) report an anonymised case study in which weekly exercise sheets “did not seem a very good predictor of final marks” with “high marks” leading to “concerns about copying” (p. 51). Cox says this means that coursework cannot provide “reliable results” (p. 151). Thomlinson, Robinson and Challis (2010a) highlight copying as a particular problem among weaker students, causing one interviewee to say that this means “it's not clear what the real benefit is of the course work” (p. 124). Iannone and Simpson (2012a) say that “some departments are moving away from the credit bearing weekly homework sheet towards in-class tests or reconfiguring them as formative assessments” because of plagiarism (p. 12).

Cox says that if a student obtains marks by “verbatim copying” then “they will be found out in the exam”. This means, he says, “the only way to make such coursework assessment harmless in this respect is to keep it at a low proportion of the overall assessment” (p. 182).

Cox says that coursework is “a good vehicle” to “encourage students to work together” provided they are encouraged to “present their own work in the end” (p. 182).

Cox says that coursework, unlike an examination, is an opportunity to give detailed feedback to students on their work (p. 181). Thomlinson, Robinson and Challis (2010a) say that coursework with rapid turnaround (“within the week”) means that staff are aware “how their students were performing” and could provide “more meaningful” feedback (p. 123). Iannone and Simpson (2012b) report on an anonymous assessment case study in which frequent, short assessments were found to have “a positive effect on final exam marks” (p. 52), apparently through helping students identify “gaps in their knowledge and misunderstandings of the material”. They report a disadvantage being increased staff marking time (p. 53).

Thomlinson, Robinson and Challis (2010a) say that “a smaller number of longer assignments encourages reflection about the interrelation of different parts of the course material and gives practice in the various stages of applying mathematics” (p. 122). Thomlinson, Robinson and Challis (2010b) say that the problem of copying “can be ameliorated to a large degree by the use of substantial open ended assignments” because these give students the opportunity to exhibit “the capacity for original thought” (p. 128). Cox says that open-ended questions “bring out how the student processes, analyses information and then synthesizes it into new forms” (p. 148).

The consistency of markers is reported as a problem by Thomlinson, Robinson and Challis (2010b), particularly for open-ended assignments which “are not amenable to producing solutions in a highly standardised form” (p. 125).

Iannone and Simpson (2012a) report use of more in-depth assignments, including for modules which focus on problem solving (p. 10) and employment skills (p. 11).

2.3.3.4 Presentations

Challis, Houston and Stirling (2004b) say that presentations might be individual or in groups, and sometimes use peer-assessment (p. 10). They say that “often” presentations “really measure” students’ “ability to talk fluently about a topic”, “to answer questions”, to “identify and explain the key points in a complicated topic” and “skills at preparing visual aids” (p. 11).

Iannone and Simpson (2012b) report on an assessment case study in which group presentations are used during an advanced pure mathematics project (p. 33). The

lecturer reported a perception of difficulty over the reliability of the marks for presentations, and that these were kept low (10% of the module mark) accordingly (p. 34).

Iannone and Simpson (2012a) report use of presentations, including in mathematical modelling (p. 11).

2.3.3.5 Peer assessment

Challis, Houston and Stirling (2004a) say that in peer assessment, students “are making judgements about their fellow students” in order that they “learn to give feedback and to give constructive criticism”. They say peer assessment can be used in “many situations in MSOR”, usually for formative assessment, but that if it is to be used for “high stakes assessment” then “students should be well trained in whatever assessment they are using” (p. 27).

Iannone and Simpson (2012a) report the abandonment of peer involvement in assessment at one university “after students raised concerns about whether peers were qualified to make robust judgements” (p. 13).

2.3.3.6 Posters

Iannone and Simpson (2012b) report on a case study which uses assessment by poster. Students were asked to create an individual poster and present it. Reported advantages include making deeper connections across mathematics and its applications, and improving independent working and oral and written presentation skills (p. 39). They raise concerns over “fairness and equality” around students becoming “extremely nervous” over the presentation of the poster (p. 40).

2.3.3.7 Oral assessment

Challis, Houston and Stirling (2004b) list “oral exams” as “face-to-face interviews with one (or usually more) examiner” which are “sometimes used to investigate borderline candidates on degree classifications” (p. 10). They say that “often” this kind of interview “really assesses” student understanding of what they wrote in an assessment, and “an overall understanding of an area, rather than a narrow question”

(p. 11). This kind of oral interview might be used in projects (Challis, Houston and Stirling, 2004a; p. 46), for example a case study reported by Iannone and Simpson (2012b) uses a panel interview (p. 49).

Oral assessment can, however, go further than this. Iannone and Simpson (2012c) focus their attention on what they call “oral performance assessment”, which “requires students to work live on a problem with a tutor” (p. 128). Cox (2011) says that oral examinations are used “in Europe and Russia”, with the advantage that staff “almost certainly get a much better idea of what a student really knows”, but notes that this is “simply too (wo)manpower intensive and beyond the resources of most UK departments with current student numbers” (p. 147). Indeed, Cox reports that written exams were originally brought in to replace oral assessment when the latter became “too costly” in the face of increased student numbers (p. 150).

Iannone and Simpson (2012c) believe this approach is “less likely to enable students to gain high marks without understanding and was more likely to engage the students in thinking about the material”. Disadvantages include student anxiety. They report that staff workload was “broadly similar to the ordinary coursework/tutorial system” (p. 134).

2.3.3.8 A note about group work

Challis, Houston and Stirling (2004a) say that working with others is “an important key skill” and “a useful learning experience” (p. 46). Nevertheless, group work is unusual in this list of assessment methods because it is not necessarily a method of assessment itself. Group work is a learning and teaching activity that ultimately produces some output, such as a report, poster or presentation, and might be assessed by methods such as oral interview or peer assessment. Group work may use many of the assessment methods discussed in this section, but it is not included here as a method of assessment itself.

Challis, Houston and Stirling (2004a) allow “group assessment” to include “assessment of the *products* of the working together of a group by either the lecturer or student peers, and to the assessment of the group-working *process*, again by the lecturer or by student peers” (p. 27; emphasis original). They say the product “will

usually be a written report or a seminar presentation or a poster. It should demonstrate the group's success in carrying out the investigation and in reporting it". They suggest marking a written report "as if it had come from one person and, initially, to give that mark to each member of the group" (p. 28).

The group-working process might include how the group arranges its activities. Iannone and Simpson (2012a) note "an interesting tension" in their interviews regarding group work, where "some described this as an important employability skill which a degree should provide, while others worried about how to measure a student's individual contribution" (p. 12). Iannone and Simpson (2012b) note the effect within groups of variability of "commitment and abilities" (p. 36), and concern about the award of "a single mark for everyone in the team when [students] perceive that not everyone contributed equally" (p. 42). Challis, Houston and Stirling (2004a) refer to "the rogue students who do not pull their weight in the process and who are carried by the others in the group" (p. 29). (MacBean, Graham and Sangwin, 2001, call these students "passengers".)

Challis, Houston and Stirling (2004a) say that "confidential self and peer assessment of the group working process can be used to good effect to moderate the mark assigned to the group in order to give a mark to each individual" (p. 29). They say this peer assessment "often gives the best insights" when assessing the group-working process (p. 27). As an alternative, they suggest the lecturer could observe the group at work, though they say "there is a danger that the process of observation will distort that which is being observed" (p. 46). (See also the discussion of peer assessment in section 2.3.3.5.)

Iannone and Simpson (2012a) report use of group projects in mathematical modelling modules (p. 11).

2.4 Systematic review of literature on advantages and limitations of e-assessment

2.4.1 Introduction

Having considered assessment in general in the previous section, here an attempt is made to understand the practicalities, advantages and limitations of e-assessment technology in the MSOR context via a systematic review of relevant literature.

2.4.2 Method

Online searches were conducted including the following research databases: Academic Search Complete; British Education Index; Educational Resources Information Center; Intute: Science, Engineering, & Technology; JSTOR; Science Direct; Web of Knowledge; Zetoc. Six searches were conducted using the search term 'mathematics' with either 'university' or 'higher education' and one of 'e-assessment', 'computer aided assessment' or 'computer assisted assessment'. In addition, searches were completed using Google Scholar with the search term 'mathematics' and one of 'e-assessment', 'computer aided assessment' or 'computer assisted assessment' in the title of the article. Finally, several professional resources were consulted: the Higher Education Academy's *MSOR Connections* archive (published approximately termly since 2000) with searches for 'e-assessment', 'computer aided assessment' and 'computer assisted assessment'; the literature library database compiled by the MU-MAP Project (University of East Anglia, 2012) with searches for 'e-assessment', 'computer aided assessment' and 'computer assisted assessment'; and, every article in the Higher Education Academy's *Maths-CAA Series* (published monthly from July 2001 and June 2006) was inspected. These searches were completed in September and October 2012.

622 articles were screened for duplicates and irrelevance so any paper not directly concerned with e-assessment in mathematics at university level was excluded. Excluded articles were most often concerned with pre-university education or with

computer-based instruction rather than assessment. Among the professional sources, some articles were simply reports of workshops or accounts of technical implementation (details of standards, etc.) and these were excluded unless they contained additional commentary from the author. 97 articles were read and notes taken where these discussed the practicalities, advantages and limitations of different aspects of and approaches to e-assessment. These notes were organised into themes, and these themes were used to write the sections of this review.

In discussing the merits of e-assessment, Foster (2004) assumes that his e-assessment tool has the functionality it claims, is reliable and stable and is correctly implemented (p. 6). This approach is taken here as well, to ignore issues of poor implementation and technical difficulties that involve software not working as advertised. In addition, issues of local support (financial, technical and political) are ignored (though they are a great cause of concern for Foster). Obviously, such issues are of interest to those trying to implement e-assessment, but this review is concerned with the principles of the advantages and limitations of e-assessment and it would be unreasonable to mark down the whole concept because of a software bug, poor interaction with other university systems or an unsupportive head of department. Certain practical difficulties are ignored as 'of their time' and now less relevant. For example, Beevers and Paterson (2003) are concerned about the "availability of a computer" and the "finance implications of using Internet access" (p. 128), which may have been a problem ten or more years ago as reported but are less likely to be an issue today. Care was taken to distinguish between discussion of e-assessment and discussion of assessment methods generally. For example, in a section on the benefits of an approach taken by Robinson et al. (2005) the discussion is actually about the benefits of diagnostic testing and not necessarily related to e-assessment; care is taken not to claim the advantages of diagnostic testing as advantages of e-assessment.

2.4.3 Findings overview

65 articles give a project report on an implementation or development of some e-assessment system, 19 offer discussion of some issue or issues from a practitioner

point of view but not directly related to a specific project implementation, eight report a piece of research that was not simply an evaluation of student feedback on a system whose implementation is the main focus of the article, four give software reviews and one is a workshop report.

30 systems are discussed (the number of articles is given in brackets where it is greater than one): AIM (10); Mathletics (9); CALM (7); HELM (4); MapleTA (4); STACK (4); i-assess (3); WebCT (3); Questionmark/Questionmark Perception (3); CABLE (2); Calmaeth (2); CALMAT (2); DEWIS (2); eGrade (2); MaxQTeX (2); Mathwise (2); METRIC (2); TREEFROG (2); WaLLiS (2); DERIVE; Maple; MathAssess; Mathwright; Moodle; Open Mark; Respondus; TAL; TASMAT; WeBWorK; WileyPlus. In addition, fifteen papers discuss custom implementations of e-assessment developed in-house and not named or discussed elsewhere.

2.4.4 Types of e-assessment and their uses

Most papers report use for routine practice at the transition between school and university “where subject material is basic” (McCabe, Heal and White, 2001; p. 1), including to deal with increased diversity of students (Ramsden, 2004; p. 2), to build student confidence (Samuels, 2007; p. 12) or for retention (Fletcher et al., 2009; p. 3). Some use e-assessment for second year (Foster, 2007; p. 45), service teaching (Blanco and Ginovart, 2012; p. 355), diagnostic testing (Pinto et al., 2007) or mathematics support (Samuels; p. 11). Some say e-assessment can be useful for some activities at all university levels (Strickland, 2002; p. 29).

Many report e-assessment as being short questions, using either multiple-choice or numeric input, in which the mathematics completed “can only be inferred from the student response” (Golden, 2003; p. 3).

Numeric input questions are of limited use in mathematics (Sangwin, 2007; p. 987) and there is a danger of penalising harshly for a small mistake, which may cause student perception of unfairness (Green et al., 2004; p. 11).

Sangwin (2007) calls multiple-choice and similar questions “provided response questions” (p. 988) as the student is given the correct answer and several incorrect responses called “distractors” (Green et al.; p. 13) or “mal-rules” (Greenhow, 2010; p.

65). Multiple-choice questions can avoid problems of user input (Green et al.; p. 11), increase curriculum coverage (Pollock, 2002; p. 249) and build student confidence (Greenhow, 2008; p. 9).

However, multiple-choice questions may allow marks for successful guessing (Pidcock, Palipana and Green, 2004; p. 300). Increasing the number of distractors would decrease the chance of guessing correctly, but this is limited by “the difficulty of producing plausible incorrect answers” (Lawson, 2002; p. 2). Negative marking can be used, though this “actively” penalises mistakes which may have ordinarily been marked zero or given partial credit, and can substantially affect marks for marginal students who only guess a small number of answers (Lawson, 2002; p. 2).

Multiple-choice can allow students who do not know how to begin a question a hint of the form of the answer (Lawson, 2002; p. 2) and opportunity to answer by process of elimination (Sangwin, 2007; p. 988). Randomisation of values may make the answer obvious (Griffin, 2004; p. 3). In some cases, multiple-choice may not test the desired learning outcomes. For example, an integral may be answered by differentiating the response options, applying the technique of differential calculus when integral calculus was supposed to be tested (Lawson, 2002; p. 6). Sangwin (2012) says the difference between solving a problem and selecting a response means that the purpose of many questions has been “grotesquely distorted” by multiple-choice (p. 3).

Randomisation might be “random selection” (selected from a question bank) or “random generation” (randomised parameters inserted into a question template) (Gwynllyw and Henderson, 2009; p. 39).

Random selection can be an efficient way of setting assessments on a standard topic (Robinson, Hernandez-Martinez and Broughton, 2012; p. 106). However, writing a question bank can be resource-intensive (Gwynllyw and Henderson; p. 39), which may limit the number, range or variety of questions (Pitcher, 2002; p. 2). This may lead to dissatisfaction with the range or level of the questions (Robinson, Hernandez-Martinez and Broughton; pp. 106 and 114) and concerns about students seeing questions more than once if the same database is used for formative and summative attempts (Cornish, Goldie and Robinson, 2006; p. 8).

Random generation can produce very large numbers of versions of a question from a single template (Gwynllyw and Henderson; p. 39), though care must be taken that these all have solutions (Greenhow, Nichols and Gill, 2003; p. 25), that all distractors and the correct answer are unique and distinct (Greenhow, 2010; p. 65), that questions are of consistent difficulty (Griffin; p. 4) and that any application context remains realistic for all random values (Jordan, Butcher and Ross, 2003; p. 7).

Randomisation enables individualised assessment. This provides the opportunity for repeated practice (Naismith and Sangwin, 2004a; p. 238), reuse of questions for future years or resits (Pitcher; p. 2) and avoidance of plagiarism (Hatt, 2007; p. 382). Plagiarism was discussed in section 2.3.2.8; here it might be taken as being simple copying from other students (Beevers, 2006; p. 4), collusion through collaborative working taken too far (Cooper, 2002; p. 2), impersonation (Beevers et al., 1999; p. 31) or something more technical like capturing screenshots during a test for a friend or repeatedly accessing a question bank to view all options (Cornish, Goldie and Robinson; p. 2). Individualisation might push students from copying answers to discussing methods, a “far more positive” peer interaction (Hermans, 2004; p. 5). Individualisation allows invigilated tests to be replaced with uninvigilated coursework (McCabe, 2009; p. 224), though this does not tackle the problem of one student completing the work for another, perhaps as a work trade (Greenhow, 2008; p. 9). Davis et al. (2005) compare results from an unsupervised computer assessment with a “similar” invigilated paper test and draw a “strong conclusion” that “cheating in CAA is not a significant problem” (p. 69), whereas Greenhow (2002) warns that “uninvigilated and repeatable-on-demand” tests “do not appear to rank students correctly” (p. 15). Regardless whether the problem is real, the potential may cause concern among lecturers (Robinson, Hernandez-Martinez and Broughton; p. 19) and students (Green et al.; p. 9).

Martin and Greenhow (2004) say that practising similar, randomised questions and examining the feedback might cause students to spot patterns and gain a deeper understanding (p. 49), though others are concerned that repeated practice on very similar questions encourages procedural ability over conceptual understanding (Robinson, Hernandez-Martinez and Broughton; p. 114). Even if we believe students

could gain a deeper understanding from studying automated feedback, this may not happen in practice; Sangwin (2003) agrees that students who adopt a deeper approach to learning are more successful, but says that students tend towards a surface approach (p. 814).

Some see provided-response e-assessment as principally a learning tool, not a “discriminating assessment tool” (Judd and Keady, 2003; p. 11), either as encouragement to practice (Sims Williams and Barry, 2003; p. 3) or for students to identify their areas of weakness (Brouwer et al., 2009; p. 282). Beevers (2000) found students more willing to test themselves than study worked examples (p. 17).

Attempts have been made to assess higher-level learning. McCabe, Heal and White suggest using “advanced question types, from ordering and hotspot to drag and drop assembly and essay” (p. 3). For proof, they suggest ordering fragments or filling in blanks in an incomplete proof (p. 6). Lawson (2002) says that such approaches are “undoubtedly imaginative use of current technology”, but “cannot be thought of as equivalent to asking a student to prove [a conjecture] from scratch” (p. 9), though such questions might help students learn “general ideas” about proofs (p. 10).

Free-text input can be collected, although marking causes problems. Simple string matching is problematic as there are usually multiple, perhaps infinitely many, ways to express the same mathematical object (Strickland, 2001; p. 3) and determining correctness may require mathematical operations (Klai, Kolokolnikov and Van den Bergh, 2000; p. 89). Some use systems which compare the student response and the correct answer evaluated at several numerical values (Beevers, 2006; p. 1), but this ignores structure and can mark incorrectly (Strickland, 2001; p. 3). A computer algebra system (CAS) can provide a powerful approach to assessing free-text input where this is available (Naismith and Sangwin, 2004b; p. 4).

CAS questions are “pedagogically more valuable than multiple choice” (Strickland, 2002; p. 28) and, though it cannot assess the whole curriculum, “a large proportion of core topics” are possible, such as “algebra, calculus, linear algebra and differential equations” (Sangwin, 2004; p. 7). Sangwin (2003) believes that “in the near future” all e-assessment will use CAS (p. 822), a prospect also reported as a future direction by Hammond (2010; p. 52).

The ability of CAS to test algebraic equivalence allows questions which require “constructing instances of mathematical objects satisfying certain properties”, a higher-level skill task (Sangwin, 2003; p. 826). Such questions are difficult in traditional assessment because marking is more labour-intensive (Sangwin, 2012; p. 14). Sangwin (2004) “firmly” believes such questions “are possible and desirable at all levels”, and have “the potential” to assess “many, if not all” topics, though this might require a question to be altered “to provide a little freedom” (p. 8).

Free text input usually requires a linear syntax, which can be difficult for students to correctly enter (Pitcher, Goldfinch and Beevers, 2002; p. 168). Since an informal linear syntax for mathematics is impossible to precisely define (Sangwin, 2012; p. 9), systems either use some non-standard syntax (Lawson, 2002; p. 5) or the syntax of an underlying CAS (Sangwin, 2007; p. 988). Pitcher calls it an “inherent problem” that “we write mathematics using symbols that do not lend themselves readily to single line input” (p. 2).

Some systems use a process that Sangwin (2012) calls “double submission”, in which input is interpreted and displayed in two-dimensional format for checking (p. 10). Strickland (2002) says that students found this “helpful” but warns that this “occasionally...misses the point or is misleading” and says that “despite the help, difficulties with syntax remained a significant cause of student dissatisfaction” (p. 28). Jones (2008) suggests using a practice question, saying that “two or three attempts” are “sufficient to allow for any typographical mistakes” (p. 354). Sangwin (2012) says that, with practice, “the majority of our students cease to find the interface especially problematic for the majority of questions” (p. 9).

Such input mechanisms add extra learning requirements (Lawson, 2002; p. 5) and cognitive load (Mavrikis and Maciocia, 2003; p. 10) unrelated to the assessment objectives. However, some regard requiring students to be precise as an advantage of e-assessment (Klai, Kolokolnikov and Van den Bergh; p. 90) and entering notation into a computer an ability students should develop (Pountney, Leinbach and Etchells, 2002; p. 33).

Some report students working on paper away from the computer even for computer-submitted work (Cooper; p. 4). Some encourage or require this, even using

the paper working to award partial credit (Pollock; p. 265). There may be pedagogic differences; Pitcher, Goldfinch and Beevers identify students who felt they would do better on e-assessment because “there was less to write” (p. 168).

In mathematics, marks are often given for the correct method with a small mistake, as this satisfies some of the learning outcomes (Lawson, 2002; p. 7). The final answer entered into a computer is not necessarily enough information to allocate partial credit (Levesley and Georgoulis, 2010; p. 1) and this can be a cause of student complaint (Croft et al., 2001; p. 65). One option for partial credit is decreasing marks for repeated submission, for example Delius (2004) penalises 10% per attempt (p. 44), though this is rather an imprecise measure and students only receive partial credit if they eventually enter a correct answer. Alternatively, a question may be broken into a series of smaller steps (Lawson, 2002; p. 7), perhaps optionally (Ashton and Youngson, 2004; p. 3). CAS can recalculate the answer for a later step as though the incorrect answer for an earlier step was the correct value, leading to enhanced follow-through marking (Sangwin, 2004; p. 7). While steps might reduce the information processing load (Beevers and Paterson, 2003; pp. 132-134) and provide intermediate feedback for increased motivation (Corbalan, Paas and Cuypers, 2010; p. 701), this may cause students to focus on individual steps while not seeing “the bigger picture”, leading to a procedural knowledge and lack of deep understanding (Quinney, 2010; p. 281). Multi-stage questions might also force students to use a method they would not have chosen (Lawson, 2002; p. 8).

2.4.5 Advantages of e-assessment

Immediate feedback is usually given as a major advantage. This is a contrast to “slow” feedback from human markers (McCabe, 2002; p. 7); negative student response to the delay in feedback being an incentive to introduce e-assessment for some (Fawcett, Foster and Youd, 2008; p. 46). Continuous feedback is linked to retention (Ekins, 2008; p. 21), and increased performance and satisfaction (Cook, Hornby and Scott, 2001; p. 1), the latter particularly when it is combined with the opportunity for immediate resubmission (Blyth and Labovic, 2009; p. 985).

Lawson (2002) says that feedback can be “as detailed as the test author wishes”

(p. 1), though this is linked to time spent and expertise in writing questions, particularly as feedback is most powerful when linked to students' misconceptions (Mavrikis and Maciocia; p. 5). There is disagreement over whether a computer will in practice provide more detailed feedback than a human marker (Mavrikis and Maciocia; p. 11) or less (Schofield and Ashton, 2005; p. 2). Greenhow (2008) provides detailed feedback and notices students entering dummy responses to access this, leading him to view the system “primarily as a learning resource” (p. 9). There is some concern that automated feedback provides “another worked example, as you find in the lecture notes”, which might “struggle to...facilitate understanding in weaker students” (Robinson, Hernandez-Martinez and Broughton; p. 114), with such students finding the feedback “more difficult to 'untangle' than the original question” (Green et al.; p. 10).

Flexibility over when and where the test is taken is an advantage (Golden; p. 3) as it can reduce student stress (Robinson, Hernandez-Martinez and Broughton; p. 115). This is advantageous to students registered on part-time courses (Samuels; p. 10) or via distance learning (Baker, 2004; p. 33). Too much freedom, though, can lead to students completing work too slowly (Pollock; p. 249) and risks removing the teacher from the process (Schofield and Ashton; p. 2).

Some students may feel “less embarrassment in giving a foolish answer when it is only the machine that 'knows'” (Lawson, 2002; p. 1), which can help reduce stress (Pitcher; p. 1). However, a key advantage widely reported for staff, but not particularly compatible with the sense of anonymity, is the opportunity for monitoring to detect “at risk students” (Quinney; p. 287), evaluate “the understanding of concepts” (Tyrrell, 2002; p. 4) and identify “areas of weakness” (Fawcett, Foster and Youd; p. 46). However, Broughton, Hernandez-Martinez and Robinson (2011) warn that monitoring data may not reflect “current level of performance”, for example if a student has omitted a question on subsequent attempts which he or she feels confident about answering (p. 2).

When marking, e-assessment avoids human error (Ferrão, 2010; p. 821) and lack of objectivity (Sangwin, 2004; p. 8), leading to fairness and reliability of testing (Beevers and Paterson, 2003; p. 128). It is important, though, to monitor the system

to make sure marking errors are not taking place (Ferrão; p. 821), which could annoy students and be unfair, particularly where partial credit by resubmission is an option (Delius; p. 46).

Time saving and staff efficiency, particularly for large numbers of students, are widely reported as advantages. Increased efficiency is attributed to reduced marking; Ferrão says that this is an “even more important” advantage than quick feedback (p. 821). Other reasons given are automated distribution of tests (Robinson, Hernandez-Martinez and Broughton; p. 112), reuse of questions (Sangwin, 2004; p. 8) and reduced administration for compiling and processing marks (Angus and Watson, 2009; p. 255). As the introduction of e-assessment is viewed as costly in terms of staff time (Quinney; p. 279), even though this may lead to a greater level of in-course assessment (Fawcett, Foster and Youd; p. 48), many make a trade between the time spent developing e-assessment and that saved on marking (McCabe, Heal and White; p. 1). However, this equation may take years to balance (Tyrrell; p. 4) and e-assessment may still require “continual updating” (Sims Williams and Barry; p. 3). There is considerable agreement among those papers which comment on the topic that the time saved will be put to good educational use, by providing improved course content (Chirwa, 2008; p. 13), greater student contact time (Pollock; p. 260) or richer feedback on hand-marked work (Sangwin, 2004; p. 8). A dissenting voice is Quinney, who refers to some wishing for “increased research productivity” (p. 279).

2.4.6 Limitations of e-assessment

Though CAS can improve matters (Sangwin, 2012; p. 3), most consider e-assessment not suitable for testing conceptual understanding, for which traditional methods are used (Robinson, Hernandez-Martinez and Broughton; p. 116). Sangwin (2012) says that mathematics is in a stronger position with regard to automated marking than “essay or more subjective artistic disciplines” because of “objective notions of *correctness* and the progressive automation of mathematical knowledge” (p. 15; emphasis original). Even so, he says that marking “extended work automatically” is impossible because we cannot “easily encode the meaning of an expression, and combine this with simple logic and automatic CAS calculations”. He says that “given

the theoretical difficulties of establishing equivalence of two expressions, establishing the validity of whole arguments automatically appears totally hopeless" (p. 14).

Beevers and Paterson (2003) say e-assessment may be too structured to test problem-solving abilities (p. 141) because students do not have "freedom to work...in a variety of ways, all equally correct" (p. 142).

Writing "reliable, valid questions" is "a difficult task, requiring expertise" (Sangwin, 2012; p. 7). Lawson (2003) reports one demonstration question provided with a software package allowing correct answers to be marked incorrect through "a failure of the author not of the package" (p. 48), demonstrating the difficulty even for, presumably, users who are familiar with the system. Errors are found in questions despite testing by staff (Foster, 2007; p. 42); several authors rely on students to notice mistakes (McCabe, 2009; p. 226).

Writing questions for e-assessment might be technically difficult, for example because of the need to write code (Strickland, 2002; p. 28) or understand the minutiae of how a CAS will handle a response (Sangwin, 2007; p. 1001). Naismith and Sangwin (2004b) even warn about instructor action "that could compromise system integrity" (p. 11).

Writing questions might also be pedagogically more difficult. Questions may need to be rewritten for automation (Pitcher; p. 3), and an author must understand what skills are being tested and what is being assumed (Greenhow, Nichols and Gill; p. 25) to avoid introducing alternative or additional learning outcomes while being "creative in finding ways round" the "limitations" of e-assessment (Lawson, 2002; pp. 4-5). Assessors must understand how randomisation affects whether questions can be answered, with Sangwin (2004) confessing to "the bitter experience of setting mathematically impossible problems" (p. 6). Second-guessing of student misconceptions is required to write multiple-choice distractors (Greenhow, Nichols and Gill; pp. 25-26), to build scenarios for multi-step questions (Martin and Greenhow; p. 52) and even for feedback in CAS-based questions (Sangwin, 2004; p. 6). Misconceptions are found by examining student work (Greenhow and Gill, 2004; p. 116) or "anticipated by an experienced teacher" (Greenhow, 2010; p. 65). However, e-assessment logs can reveal "unexpected" misconceptions (Jordan, Butcher and

Ross; p. 13), suggesting that this is a skill not usually employed in writing assessments.

Sangwin (2012) says that e-assessment does not claim to “achieve everything” (p. 3), and that it “could be used in inappropriate, or even harmful ways” (p. 12). Nevertheless, he points out, e-assessment “is routinely used by thousands of students in many settings” (p. 14). Several papers recommend using e-assessment as part of a range of methods (Judd and Keady; p. 11), perhaps using traditional assessment to make up for the limitations of e-assessment (Golden; p. 3). Hand-marked work can be an opportunity to give more in-depth feedback (Delius; p. 45) and test a wider range of skills, such as the application of techniques (Fawcett, Foster and Youd; p. 46) and “modelling, proof and interpretation skills that really do require a human marker” (Greenhow, 2008; p. 10). Combined tests are proposed as a possibility by Sangwin (2012), in which a “routine calculation within a longer proof” is “checked automatically...before the whole piece of work is submitted to an intelligent human marker” (p. 15).

3 Survey of mathematics lecturers' views of e-assessment

3.1 Introduction

Section 2.2.5 described 'the pattern of the technology enthusiast'. Implementation of e-assessment, of the sort described in the literature reviewed in section 2.4, is often driven by enthusiasts (Barton et al., 2012; p. 18). This may cause “innovator/novelty effects”, exaggerating the benefits (Sangwin, 2012; p. 12). In order to address this flaw in the literature and widen the pool of opinion that is feeding into this project in the planning stage, a survey was conducted to investigate lecturers' views of the advantages and disadvantages of traditional and electronic assessment and when these can be used effectively. This specifically looked to investigate the views of both users of e-assessment who are not represented in the literature review and non-users of e-assessment.

Seeking the views of non-users especially is potentially problematic as they may be difficult to locate and may not be motivated to engage with the topic. For this reason, a small-scale, targeted interview series (four participants) was conducted with users and non-users, none of whom are authors in the literature review in section 2.4. In order to assess the validity and relevance of the conclusions, a broader questionnaire phase was designed to confirm or reject the findings with a larger group. The method and data collected are described. Discussion of the findings, in the context of the literature reviewed in section 2.4, will take place in section 3.6.

3.2 Method

The terms 'e-assessment' (meaning assessment using computers) and 'traditional assessment' (meaning assessment through non-computer methods) were used.

Participants were asked to interpret these terms widely so 'e-assessment' represents any use of computers in assessment more advanced than the simple typesetting of reports or presentations. This research comprised two phases.

1. Semi-structured interviews with two lecturers, a user and a non-user of e-assessment, drawn from personal contacts at each of two UK universities. Universities were chosen to be dissimilar on two key measures which affect teaching: size of cohort and academic intake requirements. Interviews were conducted in July-November 2010. Questions were based on some literature findings though kept broad to avoid leading the interviewees. None of the four interviewees appear as authors in the literature survey. Themes and key quotes were drawn from transcripts via a constant comparative method (Thomas, 2009; p. 199) to inform phase 2.
2. A questionnaire delivered via the web. This used questions drawn from the themes and key quotes from phase 1 and not, say, from the literature review in section 2.4, so that literature and this survey may be usefully compared. The questionnaire was designed to be short to encourage completion. Statements were presented so that the interview findings did not always agree with the statements as given, to avoid the confusion of double-negative statements and to minimise the effect of acquiescence bias. Statements were grouped under the subheadings used in section 3.4 and within each subheading the arrangement of statements was randomised to minimise the effect of ordering. The questionnaire was piloted by seven volunteers found using Twitter and several minor wording changes were made following pilot respondents' recommendations. A new question was recommended by one pilot questionnaire participant (discussed in section 3.4.4). Questionnaire data from the pilot phase was not included in the main analysis because several of the questions were changed and one was added following the pilot.

The questionnaire was made available for four weeks in March 2011, with

requests to complete distributed via the author's blog, Twitter and an email asking contacts at 26 universities to circulate it among colleagues. Emails were also sent to the interview participants which asked them not to fill in the questionnaire, because this would mean duplication of their views, but to pass the link to colleagues. The questionnaire was not sent to any e-assessment mailing lists in an attempt to reduce the likelihood of dominance of views from those already represented in the literature review, although there is no way to know the identity of those e-assessment users who completed the questionnaire.

Data are treated as ordinal. Counts are given to indicate the level of agreement of the questionnaire respondents with the interview findings. The analysis is essentially qualitative.

3.3 Background

Of 53 undergraduate mathematics departments listed by Lionheart (2009) as having a UCAS entry requirement, University A was in the top 20 and University B in the bottom 20 on both academic intake requirement and size of undergraduate population. Participants A1 and A2 agreed with the description of University A as: 'a relatively large cohort of students with relatively strong academic backgrounds'. B2 agreed with the description of University B as: 'a relatively small cohort with relatively weak academic backgrounds'. B1 agreed with 'relatively weak academic backgrounds' but did not agree or disagree with 'small cohort of students,' explaining, "I'm not too sure what the typical cohort size is nowadays". All four interview participants were experienced lecturers, though for A2 this is mostly at further education level. Table 3.1 contains further background information.

Participants represented a mix of characteristics. Both users and non-users are represented, as well as those who teach pure, applied and computational subjects, those with educational qualifications and those without. Both users and non-users of technology are included in interesting combination. A2 and B1 are both users of e-assessment who use technology otherwise in their teaching, while B2 is a non-user in both categories. A1, however, is a non-user of e-assessment but does make use of

technology otherwise in his teaching; he agreed with the description of himself as 'a technology enthusiast who doesn't assess using computers.'

Code	Gender	PhD	Education qualification	User of e-assessment	User of technology in teaching otherwise	Topic taught
A1	male	maths	no	no	yes	pure maths
A2	female	maths	PGCE	yes	yes	applied maths
B1	male	maths	no	yes	yes	numerical analysis
B2	female	maths education	PhD	no	no	pure maths

Table 3.1: Information about interview participants.

38 questionnaire responses were received. Tables 3.2 and 3.3 contain background information. In tables 3.2-3.6, the following groupings are reported: 'All', responses from all 38 respondents; *Teachers* ('T'), 32 who selected that they teach mathematics at university and one who chose 'other' and explained that they "teach Maths and Computing to non-Mathematics undergraduates"; *Teacher-users* ('T-u'), 19 from T who identified as users of e-assessment; *Teacher-non-users* ('T-n-u'), 11 from T who identified as non-users; 'UK', 26 from the UK; *Non-UK* ('N-UK'), 10 from outside the UK. Only five participants identified as supporting teaching at higher education level (see table 3.2) so, due to the small number, these are not reported as a separate group.

Statement	All
I teach mathematics at higher education level in the UK	22
I support teaching of mathematics at higher education level in the UK	4
I teach mathematics at higher education level outside the UK	9
I support teaching of mathematics at higher education level outside the UK	1
Other	2

Table 3.2: Questionnaire responses to: 'Please indicate which of the following applies to you'.

Response	All	T	UK	N-UK
User	20	19	13	6
Non-user	15	11	12	3
No answer	3	2	1	1

Table 3.3: Questionnaire responses to: 'Do you consider yourself to be a user of e-assessment in your teaching?'

3.4 Data

3.4.1 Suitability of traditional and e-assessment

All interview participants spoke about e-assessment being useful for routine work and to test recall. A2 regarded as “one of the advantages of e-assessment” the ability to test “fairly basic understanding” and “recall”, in order to “make sure the students have got the basics on which to build the higher skills”. B1 described using a tightly-focused traditional assessment to test basic recall and regarded e-assessment as a potential replacement for this. Most questionnaire respondents agreed that e-assessment is well suited to routine work and to test recall (table 3.4 i-ii).

All interview participants spoke about e-assessment being suited to assessing proof and detailed mathematical arguments. B2 felt assessment was “mostly written work and they have to do proof and they have to work out exercises and I guess that's just not practical with computers.” Most questionnaire respondents agreed e-assessment is less well suited to proof and mathematical arguments (table 3.4 iii). B1 added that e-assessment is less well suited to open-ended and project work, and most questionnaire respondents agreed (table 3.4 iv).

Response³	All	T	T-u	T-n-u	UK	N-UK
i. For routine work						
1	0	0	0	0	0	0
2	4	4	1	2	3	1
3	18	15	10	4	13	3
4	11	9	6	3	6	5
5	5	4	2	2	4	1
ii. To test recall						
1	0	0	0	0	0	0
2	2	2	2	0	1	1
3	13	10	7	3	9	3
4	18	15	6	7	13	4
5	5	5	4	1	3	2
iii. For proof and mathematical arguments						
1	20	17	8	8	15	4
2	13	11	7	3	9	3
3	5	4	4	0	2	3
4	0	0	0	0	0	0
5	0	0	0	0	0	0
iv. For open-ended and project work						
1	13	11	4	6	9	3
2	12	8	5	3	9	2
3	9	9	7	1	6	3
4	1	1	1	0	1	0
5	3	3	2	1	1	2

Table 3.4: How suitable questionnaire respondents felt traditional and e-assessment are for four areas of assessment.

3.4.2 Advantages and limitations of e-assessment

All interview participants spoke about automated marking and instant feedback as advantages. B1 said e-assessment could take away the “entire burden” of marking. A2 used feedback directed to “common mistakes” in multiple choice questions to direct students to appropriate parts of the lecture notes. She regarded this as an advantage because it helped the student be aware of “how they're doing in their mathematics” and particularly when this feedback could be delivered “at the point at which they were doing the thinking rather than two weeks later when they can't actually remember what they were doing when they got it wrong”. Most questionnaire respondents agreed that speed of marking and instant feedback were reasons to consider using e-assessment (table 3.5 i).

All interview participants said there are limits to the use of e-assessment. A1 said

3 Response key: 1. Traditional assessment is very suitable; 2. Traditional assessment is more suitable; 3. Both are equally suitable; 4. e-assessment is more suitable; 5. e-assessment is very suitable.

he could take something he “would really like to do by traditional assessment” and “chip away” at parts that were possible with e-assessment. He said that while e-assessment was suitable for “multiple choice” or “some things that are slightly more sophisticated”, it was not possible to “get at” fully written out mathematical work. A2 spoke about assessment being about helping the student “communicate as a mathematician and join the mathematics community”. In this case, she said a lot of e-assessment was about “can you test the students' recall? Can you test whether they get right answers? And can you not bother giving them feedback but just use it to produce marks to categorise them?” She recognised the efficiency but called it “a total disaster in terms of students becoming mathematicians and learning and doing anything positive”. B1 spoke about “investigative-type work, open ended, almost project-type material,” saying “all that kind of assessment, report writing... project work, group work, any kind where it's more discursive, written type material, I'd find it hard to imagine how e-assessment could work in the same way.” No questionnaire respondents agreed that they could assess everything they needed through e-assessment (table 3.5 iii).

The three interview participants who had used e-assessment said they needed to adapt their assessment for the technology. A2 described converting a test from traditional to e-assessment. She said

you could not just take the traditional coursework and stick it onto the computer, so we actually had a little working group to look at what were the skills the traditional coursework was testing, what were the important things that we wanted to test ... The questions needed to be asked in quite a different way and new things needed to be created so that the electronic assessment could test the same kind of outputs that the traditional one was doing and then the two sets of questions looked very different. The e-assessment would not have worked on paper and the paper one wouldn't have worked electronically, but the idea was that they were testing the same things.

A majority of questionnaire respondents agreed with the statement “when using e-assessment, I have to adjust my assessment to match the technology” (table 3.5 ii).

A2 said traditional assessment was “more sustainable” if other lecturers took over the module because “probably most people can design [an assessment] that the students are actually going to sit and write”. B1 said his knowledge and skills limited his ability to use e-assessment. A majority of questionnaire respondents disagreed

that lecturers in general have sufficient knowledge and skills to use e-assessment (table 3.5 iv). There were five respondents who said they did not teach but worked in support of teaching; none of these agreed with the statement that lecturers in general have sufficient knowledge and skills to use e-assessment.

A2 said assessing deeper learning was possible through e-assessment, but that the assessor needed to be “more creative” in designing this. A majority of questionnaire respondents agreed with the statement “setting an assessment to test deeper knowledge with e-assessment requires a greater level of creativity” (table 3.5 v).

Response⁴	All	T	T-u	T-n-u	UK	N-UK
i. 'speed of marking and return of feedback to the student are reasons to consider using e-assessment'						
1	1	1	1	0	1	0
2	2	2	1	1	2	0
3	2	2	1	1	2	0
4	14	11	7	3	7	6
5	18	15	9	6	13	4
ii. 'when using e-assessment, I have to adjust my assessment to match the technology'						
1	1	1	1	0	1	0
2	0	0	0	0	0	0
3	9	8	4	4	7	2
4	20	18	12	5	14	5
5	8	5	2	2	4	3
iii. 'I can assess everything I need to through e-assessment'						
1	20	16	7	8	14	5
2	15	13	9	3	11	4
3	3	3	3	0	1	1
4	0	0	0	0	0	0
5	0	0	0	0	0	0
iv. 'in general, mathematics lecturers have sufficient knowledge and skills to use e-assessment'						
1	9	6	2	3	6	2
2	14	12	7	5	9	4
3	6	5	3	2	5	1
4	4	4	4	0	2	2
5	5	5	3	1	4	1
v. 'setting an assessment to test deeper knowledge with e-assessment requires a greater level of creativity'						
1	2	1	1	0	2	0
2	1	1	0	1	1	0
3	6	4	3	1	4	1
4	19	17	12	5	15	3
5	9	8	2	4	4	5

4 Response key: 1. Strongly disagree; 2. Disagree; 3. Neither agree nor disagree; 4. Agree; 5. Strongly agree.

Response	All	T	T-u	T-n-u	UK	N-UK
vi. 'it is important for students to hand write mathematics'						
1	1	0	0	0	1	0
2	4	3	3	0	3	1
3	10	9	7	2	6	3
4	13	12	6	5	10	2
5	10	8	3	4	6	4
vii. 'difficulty inputting mathematical notation is a problem when using e-assessment'						
1	0	0	0	0	0	0
2	6	5	4	1	4	2
3	3	3	2	1	2	1
4	20	18	10	6	16	3
5	9	6	3	3	4	4
viii. 'assessing by computer input is equivalent to assessing via hand written mathematics'						
1	12	9	2	6	9	2
2	17	15	10	4	11	6
3	4	3	3	0	3	0
4	5	5	4	1	3	2
5	0	0	0	0	0	0
ix. 'hand marking is always preferable to computer marking if you have the resources'						
1	3	2	2	0	2	0
2	9	8	5	3	6	3
3	11	10	5	3	9	2
4	11	9	5	4	8	3
5	4	3	2	1	1	2
x. 'human markers can be unreliable and inconsistent'						
1	1	0	0	0	0	0
2	4	4	2	2	3	1
3	4	3	2	1	4	0
4	23	20	13	5	15	8
5	6	5	2	3	4	1
xi. 'writing a mark scheme for e-assessment requires more thought because it is not marked by a person who can be adaptable'						
1	1	1	1	0	1	0
2	3	2	2	0	1	1
3	5	4	2	2	5	0
4	19	17	8	8	13	6
5	10	8	6	1	6	3
xii. 'human markers provide a greater level of feedback than an e-assessment system'						
1	2	1	1	0	1	0
2	7	6	5	1	5	2
3	14	13	7	6	13	1
4	11	9	4	4	5	6
5	4	3	2	0	2	1

Table 3.5: Questionnaire respondents' levels of agreement with 12 statements.

3.4.3 Producing mathematics

All interview participants answered yes when asked whether it was important for students to hand write mathematics. B1 regarded being able to write out a logical

argument as part of “the value of being a mathematician” and “one of the things we go on about” when trying to recruit students. A majority of questionnaire respondents agreed it is important for students to hand write mathematics (table 3.5 vi).

A1, B1 and B2 spoke about the difficulty of inputting mathematical notation into a computer and A1, A2 and B2 all compared computer input with hand writing and said assessment by computer is not equivalent to hand written mathematics. A1 said that if you are expecting the students “to write out and express carefully a mathematical argument,” then “it's very, very hard to get the equivalent of that done by e-assessment”. A1 called it “unreasonable” to ask students to enter “a perfect proof” and “try to get a computer to decide if it was right”, saying you can assess “little bits and pieces” of the proof process but this “can't replace” having a person read the work and comment on “their rigour and their logic”. A2 spoke about the constraints put on what can be input into an e-assessment system in order to enable automated marking, saying it would be hard to design a system which could mark work where students had the freedom to write “absolutely anything”. She said this means errors in student work such as incorrect use of notation or poor structure could not be assessed using e-assessment. A majority of questionnaire respondents agreed that difficulty inputting notation is a problem (table 3.5 vii) and disagreed that assessing by computer input and hand written mathematics are equivalent (table 3.5 viii).

3.4.4 Marking

B1 said a “it doesn't take that long to mark a reasonably well-focused phase test”. Earlier in the interview, B1 had suggested this sort of test as a good candidate for e-assessment. B2 said “with a small number of students it's definitely quicker to do the marking by hand”. A1 said “if you have the manpower for marking written work it's almost always better”. He caught himself part-way through a generalisation to a stronger point, saying “I can't think of any situation when the e-assessment, okay, for me it's not obvious that anything else is better.” Questionnaire responses were split fairly evenly between those who agreed, disagreed or were neutral that “hand marking is always preferable to computer marking if you have the resources” (table 3.5 ix).

A2 highlighted the issue of quality control among multiple markers, saying that automated marking was “much more consistent, and in theory reliable”. She said that an e-assessment test had been designed to replace a hand-marked piece of coursework when problems were detected with the quality of the marking from postgraduate students. The problem was not just a subtlety of consistency and reliability, A2 said “they actually found that incorrect work was being marked as correct”. A majority of questionnaire respondents agreed that “human markers can be unreliable and inconsistent” (table 3.5 x).

A2 spoke about a problem when using e-assessment and “lots of students gave a particular wrong answer that you hadn't allowed for in your mark scheme that was perhaps worth some marks”, saying “it's hard to go back afterwards” and

you do need to learn to look for what are the possible answers going to be in order to mark them ... you do need to put the effort in beforehand to make sure that you are getting the assessment to give marks to the things you want to give marks to.

A majority of questionnaire respondents agreed that a mark scheme for e-assessment requires more thought than for a human marker (table 3.5 xi).

A pilot questionnaire respondent suggested adding a statement about “the level of feedback that can be provided with a human marker vs. e-assessment”. This had not been a theme drawn from the interview transcripts, though A1 and A2 had spoken about computers being unable to replace a human commenting on the work. Questionnaire responses were split in response to the statement “human markers provide a greater level of feedback than an e-assessment system” (table 3.5 xii).

3.4.5 Own use of computers

Interview participants were asked about their own use of computers in setting and marking assessments.

When setting work, A2 and B1 use computers to generate examples. Half of the questionnaire respondents do this as well, including among Teacher-non-users (table 3.6 i). A1, A2 and B1 use computers to check their answers when setting work. About two thirds of questionnaire respondents do this as well, though this drops to about half for Teacher-non-users (table 3.6 ii).

When marking work, B1 uses computers to check student answers. Just over half of all questionnaire respondents also use computers to check students' answers when marking, though this figure is three quarters for Teacher-users and one quarter for Teacher-non-users (table 3.6 iii). B1 and B2 use computers as a marking aid to input and keep track of marks. Around three quarters of questionnaire respondents do this as well, including among Teacher-non-users (table 3.6 iv).

Response	All	T	T-u	T-n-u	UK	N-UK
i. 'I use computers to generate examples when setting work'						
yes	19	16	9	5	13	5
no	19	16	10	6	13	5
ii. 'I use computers to check my answers when setting work'						
yes	26	23	15	6	18	7
no	12	9	4	5	8	3
iii. 'I use computers to check students' answers when marking work'						
yes	20	17	14	3	14	5
no	17	14	5	8	11	5
iv. 'I use computers to input and track marks when marking work'						
yes	28	25	15	8	20	7
no	9	7	4	3	6	3

Table 3.6: Questionnaire respondents' 'yes/no' responses to four statements about their use of computers in setting and marking work.

3.5 Different groupings

Numbers are small, so a detailed analysis is not attempted. However, the pattern of questionnaire responses for those within and outside the UK is remarkably similar. The pattern for Teacher-users and Teacher-non-users is also similar, except that in general the Teacher-non-users were more sceptical about the use of e-assessment and more positive about the use of traditional assessment than were the Teacher-users. Specifically, non-users less often preferred automated marking or felt that e-assessment could provide a greater level of feedback. This may be partly lack of familiarity with e-assessment technology and it may be part of the considered reason why they do not use e-assessment. When setting work, fewer Teacher-non-users use computers to check their answers when setting work or use computers to check student answers when marking work.

3.6 Findings of the literature review and lecturer survey on e-assessment

Sections 2.4 provided a systematic review of literature on e-assessment. The findings of the survey of lecturers' views outlined in this chapter are now discussed in context of the findings of the literature review.

In the literature, e-assessment is most often used for routine practice in the early stages of a degree programme, perhaps to build confidence or engagement. In the survey, most participants agreed that e-assessment is well suited to routine work and to test recall.

Many in the literature report e-assessment as using short questions, often multiple-choice. Multiple-choice questions can avoid issues of user input and build student confidence, particularly via instant feedback. However, multiple-choice may allow marks for guessing, provide hints for students who do not know how to attempt questions and may not in fact test the desired learning outcomes.

Questions might be selected from a question bank, perhaps at random, though the resources necessary to write suitable questions may limit the number, range or variety of questions. In-question randomisation can provide massive numbers of questions from a single template, provided these are written carefully. Such in-question randomisation allows individualisation of assessments. This can allow repeated practice, reuse of questions and help avoid plagiarism. It is likely, as discussed in section 2.2.3, that repeated practice using randomised questions will not develop deeper learning, but such methods can have their place in building confidence with basic techniques.

Some systems collect free-text entry of mathematical notation from students, often processed by a computer algebra system (CAS). Such questions avoid giving a hint and provide access to some higher-level skills, but such methods are still limited in the range that can be assessed. Survey participants mostly agreed that e-assessment is less well suited to assessing proof, detailed mathematical arguments, open-ended questions and project work. CAS-based e-assessment can have problems with user input and syntax, which may impose additional, irrelevant learning outcomes. Survey participants generally agreed that difficulty inputting notation is a problem, assessing

by computer input and hand written mathematics are not equivalent and that it is important for students to hand-write mathematics. Whatever the method, it can be difficult to provide partial credit marks for a good method where one or more small mistakes affect the final answer.

Advantages of e-assessment found in the literature include the ability to provide continuous, immediate feedback, though writing good feedback adds to the resource cost of writing questions. Most survey participants agreed that speed of marking and return of instant feedback were reasons to consider using e-assessment. There is disagreement in the literature and among survey participants over whether a computer or a human marker will, in practice, provide more detailed feedback.

Other advantages in the literature are flexibility for the student, that students may feel freer to make mistakes in front of a computer, ease of monitoring of student performance, marking without human error, bias or other inconsistency, time saving and staff efficiency. Survey participants generally agreed that human markers can be unreliable and inconsistent, but were split between those who felt that hand marking or computer marking would be preferred if resources were not limited.

There is general agreement in the literature that e-assessment is limited in the range of the syllabus that may be accessed, and is not suitable for testing conceptual understanding or extended work. In the survey, no participants said they felt they could assess everything they needed through e-assessment alone.

It is acknowledged in the literature that writing new questions for e-assessment is a more difficult task than for traditional assessment, requiring specific expertise, either technically or in terms of adapting questions for e-assessment which assess the same learning outcomes or pre-judging possible student responses to exhaustively program a mark scheme or provide useful feedback. The survey found that, in general, participants had to adjust their assessment to match the technology. Most participants agreed that lecturers in general do not have sufficient knowledge and skills to use e-assessment, that writing a mark scheme for e-assessment requires more thought than for a human marker and that setting an e-assessment to test deeper knowledge requires a greater level of creativity.

4 Discussion of assessment methods and proposal of a novel approach

4.1 Introduction

In sections 2.3 and 2.4, literature was reviewed on assessment methods, and in chapter 3 a survey was conducted of lecturers' views on e-assessment. In this chapter, desirable qualities of assessment are defined, based on the findings of the previous chapters, and the assessment methods in common use in maths, stats and OR (MSOR) are examined in terms of these qualities. Different assessment methods in particular teaching and learning contexts are capable of accessing each of the desirable qualities to a different extent. Discussion and reflection leads to the proposal of a novel assessment approach which might be used to access an unusual balance of the desirable qualities.

4.2 Assessment methods in common use

In section 2.3, several maths, stats and OR (MSOR)-specific sources on teaching, learning and assessment in general, and assessment in particular, were consulted for guidance on current assessment practice. The following methods were given as being in use in MSOR: written examination (of various sorts); frequent short coursework; longer coursework assignments (which might include project reports); presentations; peer assessment; posters; oral assessment. Several of these methods might be the

result of individual or group work. In section 2.4 and chapter 3, e-assessment in mathematics was discussed at length.

4.3 Desirable qualities of assessment

In sections 2.4 and 3, e-assessment methods were discussed, and in section 2.3 assessment methods in general were discussed. These discussions included certain advantages a particular method might have, such as being particularly well-suited to returning feedback to students quickly, consistency of marking or developing graduate skills. Section 2.3.2.9 defined five basic qualities of assessment, reliability, validity, fairness and transparency, efficiency and usability, based on literature on assessment. Here, the advantages which arose from the literature and the survey are assigned to these five basic qualities.

- Reliability, which includes the consistency of marking and the extent of risk of plagiarism.
- Validity, which includes the extent to which the whole programme can be assessed, including: the range of the syllabus; graduate skills development; deep and conceptual learning; mathematical writing; computer skills.
- Fairness and transparency, including anonymity.
- Efficiency, including: for staff, the time, effort and expertise needed to set and mark an assessment; for students, the overall burden of assessment and the time taken to train students to use or take part in an assessment method.
- Usability, including flexibility for the student taking the assessment and the extent to which: feedback can be provided quickly and in detail; engagement with learning or practising of technique is encouraged; class performance can be quickly and frequently monitored. Usability is defined in terms of different stakeholders:
 - students, who benefit from useful feedback and a mark for progression;
 - lecturers, who benefit from information on how well students understand a topic;
 - administrators, who hope for a reliable measure of student achievement;
 - postgraduate courses and employers, who are looking for a valid measure

of achievement on a range of skills.

Of course, whether an assessment method can be used to maximum effect with regards to each of the qualities discussed depends on the teaching and learning context in which it is to be used, so any discussion in these terms will not be complete without a context. However, it is possible to discuss the potential of different methods against these qualities in general terms.

It is noticeable that many of these qualities of assessment are in balance with each other. For example, if a written examination is very carefully specified, it might be possible to increase reliability with multiple markers and decrease the time taken to provide feedback to students, but at the expense of limiting the depth and range of the curriculum that can be covered. Taking this further, an assessment could be extremely well-specified and made into an e-assessment, which would allow instant feedback by further limiting the range and depth. Alternatively, the assessment might be loosened up and questions made more open-ended, so this can be delivered as a piece of coursework, which might better assess depth and contribute to the development of graduate skills, but would decrease reliability and invite concern over plagiarism.

4.4 Discussion of assessment methods in terms of the desirable qualities

4.4.1 Introduction

The assessment methods in section 4.2 are discussed in the context of the qualities outlined in section 4.3. This discussion draws on the literature reviewed in sections 2.3 and 2.4, the findings of the survey in chapter 3, and personal experience.

Discussion of assessment methods without setting them in a teaching and learning context is, of course, somewhat artificial. Nevertheless, there are some principles which affect each of the qualities and these will be discussed in general terms. The aim is to explore the balance of qualities in search of an unusual arrangement which might not be accessed by current assessment methods, but which might be accessed by a novel approach.

4.4.2 Written examination

Written examination is thought to be a very reliable method. The use of timed exam conditions virtually removes the risk of plagiarism and means the questions are usually short and well-defined, allowing well-specified mark schemes which increase consistency of marking.

This method also scores very well for efficiency, since the well-defined short questions with well-specified mark schemes generally allow relatively quick marking. A fair examination will assess students on what has been learned and practised in the taught content, meaning that setting questions ought to be very straightforward for the lecturer who has already developed or delivered the module content, and taking the assessment should require no additional training for the student.

With high reliability and the possibility for anonymous marking, this method has a high level of fairness. The use of short, well-defined questions means that we can expect that students are clear about what is required of them, meaning transparency is high.

The use of short questions in a timed setting means that the validity of a written examination may be limited. It may not be possible to set questions which probe deeper understanding or measure more conceptual learning, or if this is attempted it might reduce reliability and efficiency. The extent of syllabus coverage, limited by denying students access to any resources in a closed book examination, can be extended by using an open book examination, but still the timed conditions will limit this. Graduate skills, for example writing in depth or giving presentations, are not well assessed by the method.

In terms of usability, this depends on the different stakeholders. Students might find the level of feedback is low from an examination. In principle, the amount of feedback produced could be extended beyond the bare minimum that is usually completed for a written examination, and a mid-term class test might indeed do this. However, the short and well-defined nature of the questions might mean there is little feedback to offer beyond pointing out where a mistake has been made. Students will benefit from the provision of marks for progression and award, a mark which is also useful and unambiguous for the administration. The lecturer will examine by hand

work from each student, meaning rich information about the progress of students can be obtained, albeit with reduced validity. Postgraduate courses and employers might be pleased to have a reliable measure of students' abilities on some topic, but again issues of validity have an impact on the usefulness of this information.

4.4.3 Frequent short coursework

In many ways, a short piece of coursework, set using short problem questions, is not very different from a written examination. It can be an efficient method of assessment with a high level of fairness and transparency. It might be more valid than even an open book examination because the time limit has been removed or, rather, extended so as to be practically removed, with potentially days rather than hours to complete a piece of work. However, reliability is negatively affected by increased risk of plagiarism because this coursework is not completed in supervised conditions. The usability profile is also similar to an examination, though there is usually a greater level of feedback offered to students and if coursework is frequent then class progress can be more effectively monitored by the lecturer. Such frequency reduces efficiency by increasing the marking load.

4.4.4 Longer coursework assignments

A longer coursework assignment, which might be the report of a project, can offer a high level of validity by allowing assessment of the greatest range, depth and connections across the syllabus of any method discussed. In general, this method of assessment does not just examine students' abilities, but can be used to develop learning while undertaking the assessment, particularly of graduate skills.

A loosely-defined coursework assignment might offer reduced reliability, since a less well-defined task can be open to more subjective marking. A more open-ended project is probably less likely to be at risk of plagiarism, since each student will produce a very different piece of work.

In terms of efficiency, this method requires questions to be set which allow students to explore a topic in different directions, which is more difficult. The work produced by students is likely to be longer and require close reading and reflection to

mark, causing a greater marking load.

More loosely-defined questions might mean that students are less clear on what is expected of them, impacting fairness and transparency. It is likely that an open-ended project means that students will require supervision and guidance, reducing the likelihood that anonymous marking can be offered.

Open-ended and loosely-defined questions should mean that a greater level of feedback is offered to students in order to explain the mark given, which impacts efficiency but should increase usability for students. Administrators might find that marking is less reliable. However, this kind of assessment ought to provide greater validity, increasing usability for postgraduate courses and employers.

4.4.5 Presentations

The need for an audience to be in attendance usually means that presentations are made in front of the peer group. For well-specified presentations on the same topic, this could invite plagiarism, intended or unintended, as each presenter will have seen all previous presentations. For a loosely-defined presentation in which students are expected to present different topics, the limitations in longer coursework assignments apply.

Giving presentations is a graduate skill, and presentations can be used to assess understanding of links across the syllabus, meaning this can be a more valid assessment method than, say, a written examination, though the depth is limited somewhat by the time restraint. Similarly to longer coursework assignments, assessment via presentation not only judges students' ability to give presentations, but also contributes to their development of this skill.

Marking is performance-based and can therefore be subjective, affecting reliability. Anonymity is not practical, and students may be unclear what is expected of a good presentation, affecting fairness and validity, or may require training on this, impacting efficiency. Marking is generally real time, perhaps with extra time for consideration of performance or review of a video after the presentation, making this generally an inefficient method, although group presentations can improve efficiency.

Students can be given detailed feedback and lecturers examine each performance,

meaning usability for each can be high. Administrators might find that marking is less reliable. The increased validity ought to increase usability for postgraduate courses and employers.

4.4.6 Peer assessment

Peer assessment can have a high level of validity, since it accesses the skill of providing feedback on the work of another, which is a valuable graduate skill which is not really accessed otherwise. Efficiency can be very high for staff, since much of the detailed work is done by students themselves. However, depending on the implementation, the lecturer might also mark the work to provide a reference for increased reliability. Fairness and reliability can be a real problem, since the students are not experienced markers and there is no certainty that they are assessing the work of their peers on its merits only. To increase fairness and reliability, students might be trained to complete peer assessment, decreasing efficiency. Usability depends somewhat on the implementation, since in principle students ought to benefit from more detailed feedback from a peer, but in fact they might receive disinterested feedback with which they disagree. For other stakeholders, the increased validity might be a boost, but the reduction in reliability seriously limits the effectiveness of this method.

4.4.7 Posters

In many ways, a poster is a combination of a written piece of work like coursework usually with an oral interrogation like an oral interview. The poster is likely to have less depth than a lengthy report but should still access many of the advantages and limitations of that method, while the oral interrogation allows for instant feedback, with similar advantages and limitations to oral assessment. If a poster is prepared but not presented or defended then this method is effectively a piece of coursework with the added requirement of graphic design and a restriction on length.

4.4.8 Oral assessment

In terms of ensuring that a student's work is their own, a performance can be even

more reliable than a written examination. An interview might be used to increase the reliability of a piece of coursework, by verifying that the student knows the work they claim to have completed. Since students can be prompted interactively with needed information while completing the work, this method can allow a greater range of the syllabus to be covered than a written examination, increasing validity.

As each student must be met individually, this could be an inefficient process, unless the meetings are kept as short as marking an exam script, in which case a reduced time scale would affect validity. Anonymity is not practical, and students might be unclear about what will happen during the assessment, meaning fairness and transparency might be adversely affected. Usability is high, partly because of the high reliability and validity and the high level of individual feedback and staff monitoring of student performance.

4.4.9 E-assessment

It is easy to see why some are enthusiastic about e-assessment. Generally, this method is of high reliability, particularly compared to multiple human markers who might be unreliable and inconsistent. Automated marking causes questions to be well-specified and allows perception of anonymity, which should increase fairness and transparency. Automated marking and administration of marks also makes this highly efficient to run.

Setting assessments can also be very efficient if questions are chosen from a bank, the initial development of which we are content to ignore. Feedback can be instant, meaning this can be a very useful method for students. Automatic marking and collection of marks can enable whole class monitoring, meaning this method has high usability also for staff and administrators.

However, e-assessment has reduced validity, since the range and depth that can be examined is limited by the capabilities of automated marking. The reduced validity adversely affects usability for postgraduate courses and employers. Range and depth might be improved by using CAS-based e-assessment, though this makes questions harder to set and means students must be trained in using the syntax, reducing efficiency.

4.5 Proposal for a partially-automated approach to the assessment of mathematics in higher education

Clearly not all of the desirable qualities can, practically, be maximised together in one assessment method. In fact, some aspects of the listed qualities are not at all compatible. We have seen how an overly-specified written examination might increase reliability at the expense of validity. It is not practical to assess the graduate skill of giving a presentation while offering anonymity. Instant feedback on extended work is not currently practical since a human marker will need time to review this and computers cannot currently mark such flexible work.

Each method considered, applied in a particular learning and teaching context, can be viewed as a balance of these qualities. For example, if a context requires highly reliable assessment of student ability to apply some technique, then a written examination or e-assessment is indicated; if oral communication skills are to be developed, then a presentation, poster or oral interview is indicated.

This way of thinking about assessment as a balance of desirable qualities shows that there is value to different approaches in different contexts. For example, CAS-based e-assessment is not better or worse than a multiple-choice question bank in general. The former is more suitable to increased validity at the cost of decreased efficiency, and the latter for quick, reliable practice and confidence-building. One or the other will be more suitable for particular circumstances, and neither is 'better' than the other.

The purpose of the background investigations in chapters 2 and 3 and the discussion in this chapter was to identify an opportunity for a new development. It is noticeable that the potential for randomisation outstrips the capabilities of automated marking. It is clearly possible, then, to generate randomised work that cannot be marked by computer, and this is one of the dangers of unthinking randomisation. If done deliberately, however, this could be used to generate worksheets that are individualised through in-question randomisation for which students complete their answers on paper for marking by hand. Many ordinary e-assessment approaches

make use of in-question randomisation for questions which are answered online and marked automatically; and this automatic marking means these tend to focus on mathematical techniques and algebraic equivalence. The difference in the approach proposed here is that the removal of computer input and automatic marking allows questions to be asked which, as they are marked by hand, can be directed to higher-order skills. This approach was not found in use for mathematics in the literature reviewed in previous sections, so this is claimed as a novel approach in higher education mathematics.

Such an assessment would experience reduced efficiency of marking, since different answers would be used for each student. However, this method might be used to increase reliability of coursework, via reduction of risk of plagiarism, and so increase validity by allowing some timed assessment tasks to be set as more valid pieces of coursework. Much of the added complexity and reduced efficiency in setting questions for e-assessment arises from the need to understand and second-guess an automated marking system, so removing this requirement should make setting questions more efficient than for e-assessment.

This partially-automated approach might be thought of as a hybrid between e-assessment and traditional coursework, in that e-assessment generally provides access to individualisation but with the limitation of automated marking, while coursework allows assessment of greater range and depth but incurs greater risk of plagiarism. Partially-automated, individualised coursework might allow a different balance of these desirable qualities.

In general, given the way assessment methods are viewed in this chapter, a piece of coursework can be seen as offering high validity with reduced reliability due to the increased risk of plagiarism. The reliability can be improved by either replacing the coursework with a test taken under examination conditions or by using e-assessment with in-question randomisation. Both of these reduce the validity of the assessment for reasons discussed above. A piece of work which is individualised by this partially-automated approach can be as valid as a piece of coursework, since its validity is not limited by the restrictions of automated marking. It should also increase reliability at the cost of reduced efficiency, particularly of marking. This is an

unusual balance of desirable qualities of assessment.

Since originally making this proposal, I have become aware of two approaches with similarities to the novel approach proposed, although both differ in significant ways. Both are used outside of conventional mathematics contexts and would not be suitable for use within it.

The first arose when I was asked to teach part of a computational methods module, although as I was not module leader I was not involved in the process of choosing or implementing the assessment approach. Students were given coursework to solve using, in this case, MATLAB, in which questions are slightly different for each student because of a randomised element which the students must compute using MATLAB as part of their answer. This is different to the proposed approach because the randomisation is a collaboration between the question author, who must carefully specify the questions, and the student, who in practice must actually vary the questions themselves. It is acceptable in this case because students are demonstrating their ability to meet learning outcomes around using MATLAB, but this approach would not work outside of a computing module because of the imposition of additional learning outcomes. The marking is in some ways similar to the proposed approach, in that this is completed by a human marker and so this approach could allow questions that cannot be automatically marked.

The second approach is in statistics and is described by Bidgood (2006) and Hunt (2007). An approach called 'Individualised Statistics Coursework Using Spreadsheets' (ISCUS) allows "individualised datasets to be produced, together with assignments (and solutions) that can be generated from the datasets"; the primary motivation being the reduction of plagiarism within the student group (Bidgood, p. 4). Hunt outlines the method, in which Microsoft Excel is used to generate random parameters and draw a randomised data sample from a larger data set. The student uses an individual five digit PIN as a seed which generates the data sample, and the marker uses this PIN to populate an answer sheet. A similar, more automated approach is used by Fawcett, Foster and Youd (2008) via an e-assessment system i-assess, which is later replaced by Numbas (Foster, Perfect and Youd, 2012). Here, Fawcett, Foster and Youd provide statistics assessment using "largely data response type questions

with each student being presented with a randomly generated unique dataset to analyse”, with “more descriptive parts” being assessed via multiple-choice questions (p. 46). Students entered their answers online and, though Fawcett, Foster and Youd express an intention to move towards automated marking, the system did not provide marks automatically; this “usually took about a week” (p. 47). These approaches involve individualised work which is marked by hand, so there are some similarities in that regard. However, the individualisation in these approaches is achieved by drawing a random data sample from a larger database. As such, this approach is applicable only to topics involving data.

Having decided on a partially-automated approach, remembering the discussion in section 2.2.6, it is important now to not simply press on and attempt to set individualised work in mathematics that students will take on paper and submit for marking by hand. First, we must stop and consider whether there exists a teaching and learning context in which the balance of assessment qualities offered would be more desirable than existing approaches.

5 Teaching and learning context

5.1 Introduction

In chapter 4, the suggestion was made that it might be possible to offer individualisation of work which is answered and marked manually on paper, via a partially-automated assessment method. This approach might provide an unusual combination of advantages and limitations not currently accessed in higher education mathematics, according to the literature in sections 2.3.3 and 2.4. The experience detailed in section 2.2 led to an approach where a technological intervention is not conducted merely because it is possible. If the new method is to be implemented, an educational context must be found in which it could feasibly be advantageous.

A new module was introduced to the final year of a mathematics degree programme in a UK university, aiming to develop certain graduate skills. In this chapter, previous experience of implementing a graduate skills group project and the context of the new module are described. The requirements for assessment of the new module mean that the novel approach suggested in chapter 4 will be seen to be more advantageous for one aspect of the module assessment than other approaches.

5.2 Programme and module context

5.2.1 University graduate attributes

The university at which the new module was introduced publishes a list of graduate attributes which it claims that all its students will have the opportunity to develop.

Some of these refer to development of subject knowledge, information literacy and developing practices of life-long learning, all aspects that one could argue the mathematics degree ordinarily offers. However, there were some aspects that the School felt were not addressed adequately by the current profile of modules in the final year, particularly around relating the subject to the real world, written and oral communication to a variety of audiences, reflective practice and aspects of working in a group such as leadership, responsibility and respect for others.

5.2.2 Degree programme context

The QAA Benchmark Statement (QAA, 2007) refers to “subject-specific skills”, those “knowledge and skills that are specific to areas within MSOR”, though allows considerable variety within this definition depending on the precise focus of the degree programme (p. 12).

As well as these subject-specific skills, the Benchmark Statement outlines a number of general skills which “graduates from the MSOR area will have acquired”. These are: problem-solving; taking knowledge from one area and applying it in another; general study skills, “particularly including the ability to learn independently, using a variety of media that might include books, learned journals, the internet and so on”; independent working; time management and organisation; to be adaptable, “in particular displaying readiness to address new problems from new areas”; numeracy; IT skills; and, general communication skills, “typically including the ability to work in teams, to contribute to discussions, to write coherently and to communicate results clearly” (p. 16).

The Benchmark Statement requires that a student who has reached the 'typical standard' of achievement, that is “the standard expected of the majority of honours graduates” (QAA, 2006; p. 4), should be able to:

- “demonstrate a reasonable understanding of the main body of knowledge for the programme of study”;
- “demonstrate a good level of skill in calculation and manipulation of the material within this body of knowledge”;
- “apply a range of concepts and principles in loosely-defined contexts, showing

effective judgement in the selection and application of tools and techniques”;

- “develop and evaluate logical arguments”;
- “demonstrate skill in abstracting the essentials of problems, formulating them mathematically and obtaining solutions by appropriate methods”;
- “present arguments and conclusions effectively and accurately”;
- “demonstrate appropriate general skills”;
- “demonstrate the ability to work professionally with a degree of independence, seeking assistance when needed” (QAA, 2007; p. 21).

The Statement notes that students “should meet this standard in an overall sense, not necessarily in respect of each and every of the statements listed” (p. 21).

The university prospectus and course documents are not very specific about the knowledge and skills developed by the degree programme. The former highlights that the mathematics degree programme provides a combination of knowledge and skills which is highly suited to demand from employers in many sectors.

5.2.3 Module aims and objectives

A document was prepared which detailed the aims, intended learning outcomes and teaching and assessment methods for the proposed module. I wrote this in collaboration with others in the School and it was approved through School- and University-level committees.

The module was a 15 credit, one semester, optional final year module for mathematics and joint honours students. It had no pre-, post- or co-requisites.

The module document explicitly stated that the module intended to develop skills which are needed in employment or when undertaking a research degree but which “may not be developed by traditional mathematics teaching”. These were listed as “working in depth on a problem over an extended period, writing reports, communicating mathematical results to different audiences, and working in collaboration with others” as well as articulation of graduate skills. This document also specifically referred to the university's graduate attributes scheme, saying that the module completed the School's “module profile” with graduate skills modules now at each undergraduate level.

The intended learning outcomes were that the module would develop:

1. enhanced problem-solving skills, including the ability to apply mathematical knowledge in real-world scenarios;
2. the ability to work in-depth on a problem over an extended period of time;
3. report writing skills;
4. oral presentation skills;
5. ability to communicate results using different methods;
6. ability to communicate results to audiences of differing mathematical abilities;
7. enhanced team working skills;
8. an appreciation of how groups operate;
9. ability to articulate graduate skills.

The module is activity-led. This means the learning is accomplished through student-led activities rather than through, say, delivery of lectures. This makes the assessment tasks particularly important as completing these is what drives the learning and skills development, rather than the assessment being simply a test of whether learning and skills development has taken place.

5.3 Previous experience of group work

5.3.1 Teaching and learning context

In section 2.2.4, the teaching of a graduate skills module was outlined⁵. Although it did not make use of the novel assessment approach proposed in chapter 4, the experience of running a group project in that module is directly relevant to the context in which I will propose to introduce that partially-automated assessment approach, and in fact the same project topic is used.

The module was a 20 credit, year long, compulsory second year module for mathematics students. It had no pre-, post- or co-requisites. One semester of the module was taught as statistics and the other as mathematics. The project to be discussed here is one of two in the mathematics semester. The module featured much

⁵ Reflection on teaching of that module, and in particular the student feedback questionnaire discussed in this section, was included in the submission for my Postgraduate Certificate in Higher Education.

self-directed learning and the development of skills rather than the mastery of a specific syllabus. On the 'Module Leader Report Form' for the previous year the then module leader had described the module as "a non-traditional, non-didactic module where students are able to develop generic as well as subject specific skills".

The module specification has the following stated aims:

This module will extend and develop your skills and experience in applying mathematical and statistical techniques to open-ended applied problems. You will gain further experience of working in small groups and as an individual on mathematical and statistical problems. You will learn to select and apply appropriate techniques and specialist mathematical and statistical software to help solve problems. You will gain further experience of reporting and presenting your results.

The project under discussion addressed the following intended learning outcomes:

After studying this module you should be able to:

1. Express problems in the language of mathematics.
2. Identify assumptions in a mathematical argument.
3. Use reference information from a variety of sources.
4. Communicate effectively using reports.
5. Communicate effectively using presentations.
6. Work effectively as part of a small team.
7. Effectively manage your time.

5.3.2 Module design

As well as the literature referenced in section 2.2.4 on the merits of group work, further advice was taken in designing the assessment task. That which is relevant to the main topic of this chapter is outlined here.

Challis et al. (2002) recommend teaching students how groups operate, saying "transferable skills must be taught explicitly as are all other aspects of the course: it is not sufficient to put students into groups and ask them to undertake tasks" (p. 89). The module was designed so that the group project topic was investigated by groups via self-directed learning, with formal teaching reserved for explicit skills development sessions.

There is an issue in group work of determining whether all students are contributing equally or whether the group is carrying, as MacBean, Graham and Sangwin (2001) put it, 'passengers'. They highlight the tension of assessed group work with a quote from one student: "I think that's a bit unfair, every single mark that you get is dependent on other people" (p. 7).

Hibberd (2002) recommends explicitly allocating marks for the management of the group, and suggests some peer assessment of individual contribution (p. 168).

Lowndes and Berry (2003) recommend having groups operate a committee structure which keeps “formal minutes clearly indicating actions and progress identifiable with each team member”, as well as individuals keeping a personal log and submitting a final critical evaluation of their personal and their group's working experience (p. 21).

Peer assessment of individual contribution can affect the reliability of the assessment process, as discussed in section 4.4.6. The process described by Lowndes and Berry requires that groups keep minutes and that individuals keep personal logs and prepare a critical evaluation of their working experience. In a relatively small project, I was concerned about the efficiency for the student of so many pieces of assessed work that were not directly relevant to the intended learning outcomes but were designed to improve reliability. The tool selected to attempt to ensure even contribution of team members was, therefore, group minutes.

Finally, the advice in the literature is for project work to avoid difficult new mathematics, “for a realistic expectation that students will be able to grasp... the wider skills base” (Hibberd, 2002; p. 163).

5.3.3 The topic

Art Gallery Problems were felt to be a suitable topic for a skills development project because, although it does not relate to previous study, the topic requires no prerequisite knowledge and is mathematically simple from first principles.

In their basic form, Art Gallery Problems are concerned with determining the minimum number of guards necessary for a simple polygon, the 'art gallery', to be 'guarded'. For an art gallery to be guarded, we require that every point in its interior can be connected by a straight line (line of sight) to at least one guard. Guards, then, are taken to be able to see a full 360° around themselves over any distance. For a convex polygon, one for which every point can be connected to every other point by a straight line which does not leave the polygon, it is clear that a single guard, placed anywhere, is sufficient. A convex polygon guarded by a single guard is shown in figure 5.1.

For non-convex polygons, multiple guards may be required. In the case of a gallery with a single corner, a single guard is still sufficient provided they are positioned at the corner. In this case, it is not possible to place a single guard anywhere in the polygon and retain a solution but there exists a point or points in the gallery that can be connected to every other point by line of sight. Such an arrangement is shown in figure 5.2. Naturally, galleries can be imagined which require many more guards. Figure 5.3 shows a gallery for which one guard is insufficient.

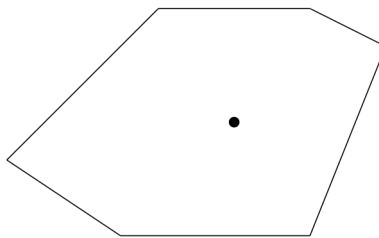


Figure 5.1: Convex polygon guarded by one guard.

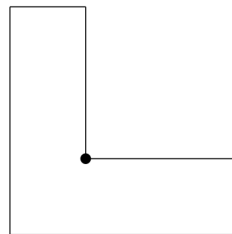


Figure 5.2: L-shaped gallery with a single guard.

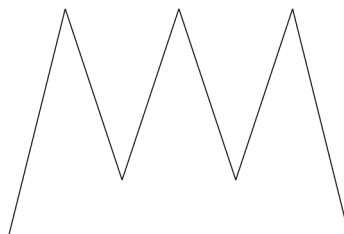


Figure 5.3: Gallery which requires more than one guard.

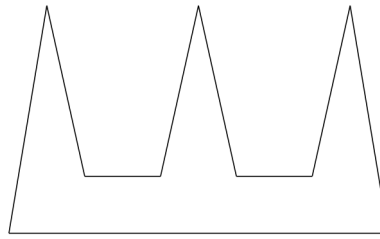


Figure 5.4: Gallery which requires $\left\lfloor \frac{n}{3} \right\rfloor$ guards, where n is the number of vertices.

In 1975, Chvátal published a theorem proving that $\left\lfloor \frac{n}{3} \right\rfloor$ guards are always sufficient and sometimes necessary to guard an n sided simple polygon (O'Rourke, 1987; p. 1). Figure 5.4 shows a gallery for which $\left\lfloor \frac{n}{3} \right\rfloor$ guards are necessary.

In 1978, Fisk published an alternative, simpler proof of Chvátal's Art Gallery Theorem (O'Rourke; pp. 4-5). The first step is to triangulate the polygon by adding interior diagonals. Any polygon can be triangulated (although not necessarily uniquely) (O'Rourke; p. 12). This triangulated polygon can then be three-coloured, that is coloured using not more than three colours such that no adjacent pair of vertices are the same colour (O'Rourke; p. 13). Figures 5.5 and 5.6 show the galleries from figures 5.3 and 5.4, respectively, triangulated and coloured using red, yellow and blue. As at most three colours have been used, at least one colour has been used at most $\frac{1}{3}$ of the time, i.e. for at most $\left\lfloor \frac{n}{3} \right\rfloor$ vertices. Since each triangle in the triangulation is a convex polygon, placing a guard at any point in each allows the entire interior to be guarded. Placing guards at the points coloured with the colour used least often allows every point in the polygon to be in line of sight of a guard, using at most $\left\lfloor \frac{n}{3} \right\rfloor$ guards. In figure 5.5, blue is used three times and red and yellow are used twice, meaning that guards should be placed at either the red or yellow vertices (but not a mixture). In figure 5.6, each colour is used three times, meaning that guards should be placed at all of either the red, yellow or blue vertices.

That the triangulation is not unique leads to questions of how a solution can be verified. In addition, in cases where the number of guards indicated is sufficient but

not necessary, the number must be reduced by some algorithmic process (O'Rourke; p. 27-30).

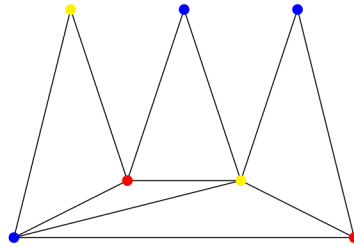


Figure 5.5: Triangulation and vertex colouring of gallery from figure 5.3.

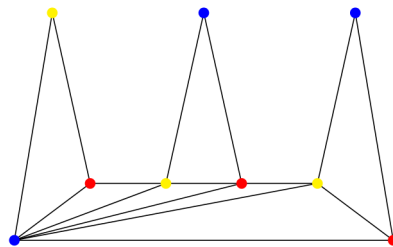


Figure 5.6: Triangulation and vertex colouring of gallery from figure 5.4.

Art Gallery Problems were introduced as “an interesting geometric problem” (O'Rourke; p. 1), and not as an applied mathematics technique. There are many simplifying assumptions made about the physical properties of the gallery and the capabilities of the guards which make this model challenging to apply in the real world. As well, there are further considerations such as where it is appropriate to place guards aesthetically, working arrangements (cover for breaks, the need to move around) and many other practical considerations which are not taken into account, but which make for an interesting student project.

5.3.4 The project

Students were asked to work in groups of three or four over an eight-week period to find the number of guards needed for each of twelve gallery floor plans and investigate an extension of the original topic of their choice. The task was framed as a problem from a client, and the students were instructed to produce a client report, a technical report and give a presentation on their chosen extension.

As the problem requires the techniques of Art Gallery Problems to be applied to a real-world scenario, this asks students to convert a scenario to a mathematical problem, identify the assumptions made (which are numerous and not subtle, so this is a good teaching exercise) and use a variety of sources to solve a problem (an initial reading list was provided), meeting the first three learning outcomes of the project. Requiring students to work in groups, manage their own affairs, prepare a report and give a group presentation on the findings, allows the latter four learning outcomes to be met.

Apart from the indicative reading list, no formal teaching was provided on the topic of Art Gallery Problems. Formal teaching was based around a series of skills sessions, aimed to give guidance on 'team work', 'questioning the assumptions in a mathematical argument', 'finding and citing references', 'report writing' and 'giving presentations'.

Ten marks out of 90 for the project were for group management. Five of these marks were for an initial plan of work, one week after the problem was set, outlining in 400 words what must be done and how it would be achieved. The groups were given a list of seven dates on which they must hold an in-class group meeting from which formal minutes were to be produced. Students were given a set of outline minutes to follow as an example. These minutes report actions and activities against group members' names and also report progress against the project plan. These minutes contributed the remaining five marks for group management.

Groups, particularly group chairs, were told that the minutes were a tool for keeping the workflow on track and for making group members accountable for their activities. The following stern warning was given with the project specification:

If a team member frequently does not attend meetings or the minutes say they are not completing work, that team member may receive a reduced share of the team mark.

Groups were invited to bring problems of uneven contribution, evidenced by minutes, to my attention.

5.3.5 Evaluation of previous group project activity

5.3.5.1 Method

A feedback questionnaire was filled out on paper in the first class after the project had been completed by 28 students, the number who were present out of the 36 total students. This analysis focuses on two aspects with direct relevance to the new module: suitability of Art Gallery Problems as a topic for group work, and the issue of uneven contribution. Recall that the topic attempted to avoid difficult mathematics in order that this did not obstruct the development of skills, and that minutes of group meetings, indicating individual accountability for actions and progress, were used to attempt to ensure even contribute and prosecute uneven contribution.

5.3.5.2 Suitability of Art Gallery Problems as a topic

Art Gallery Problems had been chosen as requiring no prerequisite knowledge and being mathematically simple from first principles. Students were asked to agree or disagree with the statement 'This project required me to use knowledge or skills gained in previous modules' and give examples if they agreed. All 28 students answered this question. Nine students agreed with the statement, but the knowledge and skills they indicated, given in table 5.1, were non-mathematical in nature. The students were also asked to rate the difficulty of the mathematical content of this project compared to other modules. The majority rated the difficulty as 'about the same' or 'easier'. The results are given in table 5.2.

Students were asked to rate their enjoyment of 'Art Gallery Problems as a topic'. 28 students answered the question and nine of these indicated that they had not enjoyed the topic. This is fewer than disliked other aspects of the project, such as 'Writing for a technical audience' (11), 'Writing for a business audience' (10) or 'Giving presentations' (12), though more than disliked 'Undertaking self-directed learning' (2) or 'Working as part of a team' (5).

Communication, team work, report writing
IT skills
Personally I used skills from previous employment rather than previous modules
Presentation and report writing skills from other modules were useful for this. Group work. Working on given problems.
Presentation skills, report writing
Presenting skills (having experience) Research skills / knowledge
Report writing skills. Knowledge of given problem. Presentation skills. Group work.
Some knowledge from GCSE work and [a year 1 module which involves skills development]
Team working skills, giving presentations

Table 5.1: Free text examples of previous knowledge or skills required for this project.

Response	Number
1 - much easier	0
2 - easier	13
3 - about the same	13
4 - more difficult	2
5 - much more difficult	0

Table 5.2: Responses to 'How would you rate the difficulty of the mathematical content of this project compared with your other modules?'

5.3.5.3 Measures to tackle uneven contribution

Groups had been invited to bring problems of uneven contribution to my attention, and these would be evidenced by the minutes of weekly team meetings. In fact, no student or group chose to do so.

Students were asked to rate how helpful they found the plan of work was in organising the project and how useful minutes of meetings were in ensuring that team members completed their assigned tasks. 26 students answered these questions. The results are in figure 5.3. This provides a positive picture, with most students reporting finding minutes of meetings to be helpful in ensuring that team members completed their assigned tasks.

Response	How helpful did you feel the plan of work was in organising the project?	How helpful did you feel the minutes of meetings were in ensuring team members completed their assigned tasks?
1 - very unhelpful or distracting	0	1
2 - quite unhelpful or distracting	0	1
3 - neither helpful nor unhelpful	6	8
4 - quite helpful	18	13
5 - very helpful	2	3

Table 5.3: Student responses to two questions about the plan of work and minutes of meetings.

Another question asked students to indicate their level of agreement with the following statement: 'All team members contributed to the project equally'. 26 students answered this question. The responses are given in table 5.4 and provide a far less positive picture. Seven students disagreed with the statement and seven more neither agreed nor disagreed. Only twelve of 26 agreed or strongly agreed that all members of their team had contributed equally.

Students had been asked to indicate whether they took a named role as Chair or Minute Taker in their team. A breakdown of the question about whether minutes of meetings were useful in ensuring team members completed their assigned activities is provided in table 5.5.

Response	Number
1 - strongly disagree	1
2 - disagree	6
3 - neither agree nor disagree	7
4 - agree	8
5 - strongly agree	4

Table 5.4: Student responses to 'All team members contributed to the project equally'.

Response	Chair	Minute taker	Ordinary group member
1 - very unhelpful or distracting	1	0	0
2 - quite unhelpful or distracting	1	0	0
3 - neither helpful nor unhelpful	0	1	7
4 - quite helpful	4	5	4
5 - very helpful	2	0	1

Table 5.5: Breakdown of responses to 'How helpful did you feel the minutes of meetings were in ensuring team members completed their assigned tasks?' by team role.

5.3.6 Discussion

Students generally agreed that the topic of Art Gallery Problems did not require prerequisite mathematical knowledge and was not more difficult than work in other modules. Most enjoyed the topic, certainly comparably to other aspects of the project. This supports the use of Art Gallery Problems as a topic. Although I only taught it once, I am aware that the Art Gallery Problems project has been set for the four academic years since I did so, further supporting the appropriateness of the topic.

The conclusions with regard to the measures for tackling uneven contribution are less positive. Fewer than half of the students reported that the members of their group had contributed evenly. Although many felt that the minutes of meetings were a useful tool for ensuring that team members completed their assigned tasks, it is noticeable that two group chairs disagreed with this. As the minutes were effectively a tool for group chairs to use to ensure even contribution, there had clearly been some failings. It seems likely that marks were awarded to students for work to which they had not contributed evenly. Putting it bluntly, there were passengers being carried by some of my groups that the 'minutes of meetings' process did not detect. One student's free text response on the questionnaire may indicate the reason for this:

Always difficult to get even shares of work amongst group and yet difficult to get 'friends' turning on each other (ie challenging ineven [sic] balance of work).

5.4 Group projects

Returning to the new module outlined in section 5.2, the main activities were completed in groups to meet the aim of team working. Three group tasks were designed to address particular intended learning outcomes, as described below.

1. A short time-scale, low stakes problem-solving task of approx. 2 hours duration. This involved finding a mathematical solution to a real-world problem (learning outcome 1) and communicating this via a report (learning outcome 3), with sections for a mathematical audience ('the boss') and non-mathematical audience ('the client') (learning outcome 6). This short-time scale, perhaps relatively high pressure task allowed each group to learn about the strengths of its members and their interactions (learning outcome 8), which was supported by a lecture, group discussion and reflection on a formative ten-minute small group exercise. In addition, this was designed to assist groups in organising themselves for more effective team working in the higher-stakes group tasks (learning outcome 7).
2. An in-depth project of approx. 3 weeks duration (learning outcome 2), in which students worked in groups (learning outcome 7) to address a specific task from a client. Groups needed to translate a given scenario into a mathematical problem, gather information from multiple sources including academic journals, and solve the problem (learning outcome 1). Having obtained a solution, groups communicated this both to other mathematicians and to the client (learning outcome 6) via reports (learning outcome 3).
3. An in-depth project of approx. 6 weeks duration (learning outcomes 2), in which students worked in groups (learning outcome 7) to propose and develop a research topic based on real-world data (learning outcome 1). Groups communicated their results both to other mathematicians and the general public (learning outcome 6) via reports, presentations and an audio media report suitable for the radio or podcast (learning outcomes 3, 4 and 5).

Given the positive previous experience described in section 5.3, group project 2 used the topic of Art Gallery Problems. The task was redesigned to involve fewer museums and to remove the open exploration of the topic, as open exploration of a

mathematical topic was saved for group project 3, and this allowed a well-specified client problem which could be solved in a relatively short time-scale group project. The opportunity to use the approach proposed in chapter 4 will occur in group project 2, so the remainder of this thesis will be concerned principally with that project.

Students completed a 3 week group project on Art Gallery Problems, presented as if for a client. Students were asked to provide justified answers for how many guards are required for each of 3 specified museum layouts. Students answered the problem in a report to a non-mathematical audience, the client. Students also documented their techniques for a mathematical audience, as an 'in-house technical report'. In this, students were asked to provide a critical analysis of the appropriateness and limitations of the model used to solve the client's problem.

5.5 Group management, individual work and uneven contribution

The previous experience, outlined in section 5.3, indicated that minutes of meetings, while a useful tool to help ensure that team members completed their activities for some, were not sufficient to ensure even contribution by all group members or to detect when uneven contribution was taking place.

Here, minutes were again used as a tool for groups to attempt to assign and monitor activities, and a group management mark was derived from these. In addition, peer assessment of contribution was used to modify the group mark for individual group members in group project 2. This does increase the workload, but this is justifiable since this time the intended learning outcomes allowed for such reflection.

As well, to further support individual marks taking account of individual contribution and comprehension of the group activities, allowance was made in the module document for individual assignments. This was in the form of three reflective essays, to develop and assess students' understanding of some of the issues in the project and to provide for learning outcome 9, and two mathematical assignments. Reflective essays were set in connection with group project 1, 'How our group operated and my part in it' (learning outcome 8), group project 2, 'What the client

wants' (learning outcome 6), and group project 3, 'The difference between mathematics at university and applying mathematics in the real world' (learning outcome 1). The individual mathematical assignments were connected to group projects 2 and 3 and required students to complete individual work on the same topic as the group project.

For group project 2, the group activity involved finding the number of guards needed to guard three client galleries. Individual assignment 2, the related mathematical assignment, required students to demonstrate that they had understood something of the mathematical topic and were able to relate this to the real-world scenario. This meant individuals needed to solve a fourth museum and discuss its solution in context of the real-world scenario. The deadline for the individual assignment was set one week ahead of the group project deadline to encourage groups to have a firm grasp of the mathematics early, in order to leave plenty of time to work on communicating the solution to the client.

The similarity of the individual and group tasks meant the risk of in-team plagiarism was high. Firstly, there is the usual risk of students who all have to complete the same piece of work copying from one another. Secondly, if students were to solve three museums with their group and a fourth, that they all have to solve, individually, the temptation to solve or partly-solve the fourth also as a group would be huge. For the remainder of this thesis, in accordance with the literature reviewed in section 2.4, the term 'plagiarism' will be used to refer to this risk and to include 'collusion'. The high risk of plagiarism here suggests that exam conditions or individualised work would be advantageous.

To suit the intended learning outcomes, the individual work required students to solve a problem and discuss its solution in the context of the real-world scenario. This deeper, more open questioning does not suit high-stress exam conditions and is beyond the limits of automated marking. In addition, hopefully it is clear from the outline of the topic in section 5.3 that solutions to questions would involve drawing diagrams. This is not well suited to computer input because it introduces additional, irrelevant learning outcomes, such as use of a drawing package.

Adaptation of the assessment to suit the available techniques therefore risks

distorting its purpose.

The requirement for randomisation, lack of suitability of automated marking and the need for students to be able to hand-write their answers suggests that this teaching and learning context precisely requires the partially-automated approach suggested in chapter 4. Any other approach, such as a standard coursework, class test or computerised test seems unlikely to meet the learning objectives to the same extent.

6 Evaluation method

6.1 Introduction

Having identified a novel approach to assessment in chapter 4, we must, drawing on the experience detailed in section 2.2, find a particular context in which the introduction of the novel approach is useful. A teaching and learning context was proposed, but not tested, in chapter 5. What remains is to implement the novel approach in the proposed context and evaluate that implementation. This chapter details the planning of the evaluation, which is put in practice during implementation in chapter 7, before its results are presented in chapter 8.

Manwaring and Calverley (1998) give a series of questions to think about when designing the framework of an evaluation study:

- “what will your evaluation do?” (p. 9);
- “who is the evaluation for?” (p. 9);
- “can you deal with the practicalities?” (p. 10);
- “what methods are best?” (p. 10);
- “what impact will it have?” (p. 10); Manwaring and Calverley mean “impact” in a negative sense, warning about “unintended consequences” and “potentially difficult situations” when dealing with the “delicate topic” of evaluation (pp. 10-11);
- “what are your deliverables?”; here, Manwaring and Calverley cover positive and deliberate impact, such as how findings will be implemented into future work (p. 11).

These areas will be considered when designing this evaluation.

6.2 What will the evaluation do?

Shaw (1998) says evaluation is “about asking questions, and then designing ways to try to find useful answers” (p. 8). Moore (2011) warns about “a tendency to try to evaluate everything about a project without due consideration as to what the central focus of the evaluation process should be” (p. 6) and says evaluation “should be focused on the core objectives or goals” (p. 8).

Most generally, the core question here is whether the proposed partially-automated assessment approach can be put to some good use in higher education mathematics. In general, assessment must “judge the extent to which knowledge and skills have been mastered” (Cox, 2011; pp. 149-150), so reliability and validity must be established for any assessment. In addition, the proposal in chapter 4 for the partially-automated assessment method suggested this might maintain the validity of a piece of coursework while increasing reliability via reduction in plagiarism, at the cost of reduced efficiency, particularly of marking. It is important, therefore, to check that reliability, with respect to who is doing the marking, and validity are not adversely affected by the use of this method, and to examine its contribution to plagiarism. Specific to this context, the individualised assessment was proposed in chapter 5 in order to reduce plagiarism in a piece of coursework as part of a package aiming to address the issue of uneven contribution in group work, so the effect of this assessment on uneven contribution should be considered.

There are, then, five pertinent questions which will be asked to evaluate the assessment in this particular context. If the implementation is successful in this case, we will have found an example of a productive use of the novel approach in higher education mathematics, satisfying the general question. These questions are:

1. Are the marks particularly sensitive to who is doing the marking?
2. Is the assignment assessing the learning outcomes it was intended to assess?
3. Does the individualised nature of the assignment work to reduce plagiarism?
4. Can the individual work contribute to a view on whether uneven contribution is taking place?

5. Can the individual work contribute to a process for adjusting group marks to take account of individual contribution?

As well as these specific goals, recall that the idea that minutes of meetings was an effective method for managing group activities was based on student feedback from one previous trial. An attempt will be made to validate or dispute that conclusion by repeating the same question in student feedback here. In addition, this group took part in a peer assessment of contribution process, and views on this will be sought. This leads to a question:

6. What are student views on conventional measures to address uneven contribution?

Finally, the justification for introducing this novel approach was based on certain conclusions about the purpose of assessment and automated marking. These views will be tested with students, leading to:

7. What are student views on the purpose of assessment?
8. What are student views on automated marking?

6.3 Who is the evaluation for?

Chelimsky (1997) identifies three “conceptual frameworks of evaluation” associated with three different “reasons for conducting evaluations” (p. 100):

- evaluation for accountability (e.g. measuring results or efficiency);
- evaluation for development (e.g. providing evaluative help to strengthen institutions); and,
- evaluation for knowledge (e.g. obtaining a deeper understanding in some specific area or policy field).

Moore says that “evaluation in the context of innovation projects is to be used as a tool for development” and should be thought of as “a formative process”. Moore identifies Chelimsky's “evaluation for development” as “providing information to help improve practice” and so appropriate for a formative process (p. 5). Thomas (2009) describes “evaluation research” in education, saying that this is “undertaken to assess how effective a programme of activity has been” in order to provide evidence for the evaluator to “decide to continue with the programme, modify it in some way or completely 'pull the plug' on the innovation – depending on the findings of the evaluation” (p. 122). Moore contrasts formative feedback, “for the purpose of

improving the learning environment”, to summative feedback, “essentially to provide evidence of the effectiveness of the environment for the benefit of others” (p. 8).

Here, the evaluation is certainly “evaluation for development”, in order to understand whether the novel assessment approach is useful in this particular context and how it might be adapted for future use. There is also an element that is “evaluation for knowledge”, both in the particular context, where we hope to learn something about evenness of contribution in group work, and in the general case, where we hope to establish whether the partially-automated assessment approach can be put to effective use in higher education mathematics. As this is a self-funded research project, we do not really have “evaluation for accountability”, in the sense of convincing an external agency that their investment has led to a successful outcome.

Shaw recommends that when planning goals, students should be considered, asking “what are the concerns of the group you are targeting?” (p. 8). The previous experience of group work detailed in section 5.3 revealed undetected uneven contribution as an issue from student feedback, so a focus on this aspect seems to be appropriate to the concerns of students.

6.4 Practicalities

Manwaring and Calverley's question here was “can you deal with the practicalities?”, and they point out that “resource levels will restrict the amount of information you can most usefully gather and process” (p. 10). Moore says that “the resource spent on evaluation should be proportional to the resource spent on the activity. Although evaluation is important, it is also important not to over-evaluate” (p. 8).

Where students are involved, it is important to consider timing. Oliver and Conole (1998) express concern about how long data gathering will take (p. 13), in a way which is reminiscent of the need to ensure a reasonable workload for students in the discussion of assessment techniques in section 2.3.2.9. The students already had to fill in a department end of module questionnaire which was not relevant to this evaluation. This is not useful because the questions were mostly suited for a traditional, didactic lecture course and not for a student-led group work module (e.g. agreement with “You could hear what the lecturer said” and “Lectures were delivered

at an acceptable speed"). In any case, the responses were not made available to me. Still, this questionnaire added to the class time and student effort that was already being devoted to evaluation, and limited further the extent of the student feedback that could be reasonably collected.

The issue of workload is also important for other volunteers who are involved. Second-marking, which will be proposed in the next section as a method for examining questions 1 and 2, relies on the kindness of volunteer markers. The amount that is asked of them must be limited to what is strictly needed.

Some of the evaluation can be conducted using the marks already compiled for the module, and this will be explored where possible to reduce the overall resource needed for collecting additional data.

The scope of the assessment is also affected by issues of practicality. Goldacre (2013a) writes in support of using randomised controlled trials in education, and praises a project (Goldacre, 2013b) which is doing so "in over 1,400 schools and with 275,000 pupils" (Education Endowment Foundation, 2013). In university teaching, the opportunity to test research questions on this scale is unusual. In reality, the chance to design and teach a new module like this is a rare privilege, and whatever evaluation can be attempted with the 44 students who took the module will have to be sufficient. Similarly, a second-marker experiment with hundreds of groups of students would produce more reliable results, but the practicalities of collecting work and having this second-marked limit the possibilities. The small numbers mean that, although some descriptive statistics and tests will be used, the analysis is for the most part essentially qualitative.

6.5 Methods

Moore says evaluation should be "convergent" and "divergent". Convergent evaluation "focuses on the intended goals" (p. 8), which Moore refers to as "goal-oriented" (p. 12). Divergent evaluation "seeks to identify unintended outcomes (goal-free) and activities beyond those of the initiative that have had an impact on goal achievement (context-free)" (p. 8). Given the practicalities discussed in the previous section, the extent to which divergent evaluation can be explored is limited,

though student feedback will make use of free-text questions where possible to explore this.

Moore also stresses that evaluation should not be pursued to “provide evidence of good practice or to find fault”, but that evaluation should be “open and honest” (p. 8). There is a famous quote about abuse of statistics to back up pre-determined conclusions, often attributed to Andrew Lang: “He uses statistics as a drunken man uses lamp-posts – for support rather than illumination” (Goldsmith, 1975, p. 58). Although this evaluation will be largely focused on the proposed questions, it will not seek to justify or find fault with any pre-determined conclusions. In particular, this evaluation will not presume that the partially-automated assessment approach works well.

The remainder of this section will discuss methods for each of the proposed evaluation questions in turn.

6.5.1 Are the marks particularly sensitive to who is doing the marking?

Question 1 examines reliability, in that it asks whether the assessment is sensitive to who is the assessor.

To some extent, evaluation of the reliability and validity will be carried out through departmental moderation and scrutiny of the outcomes of the assessment, and through the work of the external examiner (Cox; p. 146). For this module, in fact, the minutes for the examination board report that the board was told that student work had been second-marked, and that “marks were consistent, as was the ranking”. Relying on these existing processes would reduce the resource load discussed in the previous section. However, these processes are insufficient for this evaluation, not least because they work at the module level and the novel approach was used only on a small component of the module assessment. Moderation of work only took place on the main group project deliverables, not the individual work. Beyond this, Bloxham (2009) criticises the inherent assumption that internal or external moderation can ensure consistency of marking in higher education (p. 212).

A second-marker experiment would be a straightforward way to test the objectivity

and accuracy of multiple markers. We should not expect complete agreement between multiple markers for this deeper, more open-ended form of assessment. Recall that Cox says that “any [mark] scheme that is sufficiently detailed to always guarantee reliability is almost certain to be complex and unwieldy, or to render the question so anodyne as to pose little challenge” (p. 151). As well, Bloxham criticises the inherent assumptions that higher education work can be awarded an accurate and reliable mark (p. 209) and that academics share common views regarding academic standards (p. 210).

Since we cannot reasonably expect complete agreement, conclusions about whether the level of agreement found between multiple markers is reasonable or not require context. In order to calibrate expectations and provide reference information, we may examine the level of agreement for multiple markers of other, more established assessments. The difference between marks from multiple assessors for two assessments will be used for reference: a class test under examination conditions, a method of assessment recognised as being highly reliable (Cox; p. 151); and, an open-ended piece of coursework, a method reported as having problems with consistency of marking (Iannone and Simpson, 2012a; p. 12).

6.5.2 Is the assignment assessing the learning outcomes it was intended to assess?

Question 2 concerns validity. Recall that Cox said that validity was “the requirement that the assessment measures attainment of the learning objectives set” (p. 151).

Given that second-markers will be recruited and asked to examine the mark scheme and student work, a simple test of validity would be to ask the second-markers what they think the assessment is assessing and see whether this matches the intended learning outcomes.

For this reason, care will be taken to control what information is given to the second-markers. They must be shown student work and given enough information to mark this for the experiment to take place, but they will not be told the intended learning outcomes.

6.5.3 Does the individualised nature of the assignment work to reduce plagiarism?

Plagiarism might be detected through the normal marking process, in that I might notice if work is very similar when I am marking it.

Beyond this, the risk really is around intra-group plagiarism, since members of the group are working together on similar problems. Inter-group plagiarism may be a risk but this seems less likely, firstly because the groups were partly chosen by the students, meaning that they are likely to have their close friends in the same group, and secondly because of the sense of inter-group competition that exists in group work.

Individual marks from within groups can be examined for differences. If there is wide variety within groups, we might conclude that intra-group plagiarism is not a large problem. If there is not wide variety, however, this may not indicate plagiarism. Since the groups are expected, in an ideal situation, to learn about the topic and how to approach problems together, we might expect that even contribution to researching the topic could cause their level of understanding to be similar.

A third approach to detecting plagiarism could be to ask the views of students. Students will be asked in a questionnaire to express their views, anonymously, on the role of individualised work and how this affected interaction with other students, as well as general questions about plagiarism in this assignment and other university work.

6.5.4 Can the individual work contribute to a view on whether uneven contribution is taking place?

It is important to define what is meant by uneven contribution. A natural interpretation of evenness of contribution in group work might ask whether each member of the group has 'pulled their weight' and put in the same amount of effort. It seems very likely that this is the sense which students use when they fill in a peer assessment of contribution form. However, at university it is not usual to award marks on the basis of effort alone. Rather, assessment should be aligned to the intended

learning outcomes, so that a score on an assessment provides some measure of how well the student has met the intended learning outcomes (Kahn, 2002; p. 102).

The individual assignment attempted to assess a subset of the skills targeted by the group project. This could be used as a measure of how capable students are to contribute evenly to the group project. For this reason, we might expect some correlation between the individual marks and the group mark. If this correlation is not present, and there is a large dispersion of marks within each group, this would suggest that uneven contribution was taking place.

The measure is not precise. Firstly, the student may have made a valuable contribution to the same learning outcomes in the group project but have been unable to express this in the individual work for some reason. Secondly, the group project aims to assess a wider range of learning outcomes than does the individual assignment. Thirdly, Kahn points out that learning outcomes “will never fully characterize a student's understanding” of the topic (p. 93). It seems, though, a reasonable approximation to say that if a student cannot demonstrate ability at some learning outcome in the individual coursework report, then they are unlikely to have been the one who demonstrated that ability in the group project report submitted one week later.

Linked to the previous question, if there is little in-group dispersion of marks then this might indicate even contribution, or it might indicate plagiarism has taken place.

6.5.5 Can the individual work contribute to a process for adjusting group marks to take account of individual contribution?

In the module, in practice, group marks were adjusted according to the peer assessment of contribution. However, for reasons discussed above, it might be preferable to use individual work targeted at the same learning outcomes as a measure of individual ability to contribute to the main group project. When examining question 4, whether uneven contribution is taking place, the marks will also be examined to try to determine whether the individual marks can produce a useful measure of unevenness.

6.5.6 What are student views on conventional measures to address uneven contribution?

The previous experience of group work described in section 5.3 used minutes of meetings as a tool to try to address uneven contribution. The conclusion from the feedback in that trial was that students generally felt that the minutes were useful, but that uneven contribution still took place. Minutes were again used here, as a structuring device. The same question, 'How helpful did you feel the minutes of meetings were in ensuring team members completed their assigned tasks?', will be used again to validate or dispute the conclusion from the previous trial.

In addition, this group took part in a peer assessment of contribution process, and views on this process will be sought. The peer assessment itself collects views on how evenly the students felt that their peers contributed. A questionnaire will be used to explore the peer assessment through a divergent evaluation approach.

6.5.7 What are student views on the purpose of assessment?

Chapter 4 discussed assessment methods in terms of their various advantages and disadvantages. A questionnaire will encourage students to share general views on the purposes of assessment, to provide data to help explore these issues through a divergent evaluation approach.

6.5.8 What are student views on automated marking?

Chapter 5 proposed a novel assessment approach partly based on the limitations of automated marking. A questionnaire will explore students' views on the differences between computer and manual marking methods through a divergent evaluation approach.

6.6 Potential negative impact and ethics

Oliver and Conole warn about the ethics of testing ideas “on students whose exam grades may be affected”. However, they say, tests done in a “laboratory-like setting” are not “authentic” and may not produce useful information on how techniques might be “useful as part of the curriculum” (p. 12). Kounin (1970) refers to a classroom as having “its own ecology” which means that “being a pupil in a classroom is not the same as being ... a subject in an experiment. And being a teacher is different from being ... [an] experimenter”. For this reason, he warns about the need to be “cautious in applying research findings from experiments” without “validating them in the realities of classrooms”. Specifically, he warns that “variables that may be significant in an experiment may be so outweighed in importance by certain ecological variables that the former just don't make a noticeable difference in classrooms” (p. 59). Thomas picks up the idea of the classroom as an ecology, writing about problems that are encountered when trying to “manipulate social situations as tightly as is necessary for the running of an experiment” (p. 129).

To draw conclusions that are 'authentic' and 'useful', then, evaluation should take place in a live teaching situation, which raises ethical issues of impact on student grades, particularly so for a final year module. For this reason, the teaching used the better established peer assessment of contribution as the main driver of group marks adjustment for individuals. Something similar had been used in the second year of the same degree programme so this seems reasonable. Then the individual work was used to only contribute a small proportion of the overall mark (4% of the module) in order to keep the untested new approach from having an undue influence on the final grade.

To avoid unwanted impact for those involved, Manwaring and Calverley say, “some aspects may need to be confidential and anonymous” (p. 11). Here, since student marks and the process of assigning those marks are to be discussed in detail, the identity of the university and everyone involved will be kept confidential. Additionally, the second-marker experiments require further student work and marks completed by multiple assessors, and all names and institutional affiliations will be kept confidential

here as well.

6.7 Deliverables and plans for positive impact

Manwaring and Calverley ask “how will the results of the study be distributed and to whom?” and “how will the results be implemented into your work?” (p. 11). Moore says that evaluation should benefit “both the learner and the educator” by helping to (p. 5)

identify the impact of your teaching practice on the student learning experience; to provide you with information about how you might improve your own teaching practice; and hopefully to provide you with an evidence base which will help you to turn your findings of good practice into scholarly publication.

Moore provides a “model for developing professional teaching practice”, in which reflection on evaluation outcomes leads to further scholarship, which in turn forms part of the “‘theory’ of teaching that will inform and improve your practice in the future” (p. 14).

If the reader will forgive the momentary breaking of the fourth wall, I will remark at this point on the importance to the process of this project being for a PhD.

Far from simply being the report of a research project completed for my PhD, writing this thesis is part of the further scholarship that will feed into improved teaching practice, by way of evaluation. There should, therefore, be positive impact on my future students from the deeper understanding and proposals for further work discussed in chapter 9.

The importance of the thesis goes beyond this, however, because holding a PhD is central to building my own employability profile to the stage where I might hope to be appointed to an academic position and be able to continue to work in university teaching. Assuming that the outcome is positive, and that I am able to obtain further work in the higher education mathematics teaching, the opportunity exists for writing this thesis to be an initial stage in lifetime of practice.

As well as the pragmatic importance to my future activity, the scope of the evaluation is affected by the nature of the project. Moore is concerned that the “resource spent on evaluation” be “proportional to the resource spent on the activity”

(p. 8). In this case, it is rational to spend more time on evaluation than would be practical for most other aspects of my teaching because it is for my PhD thesis. Apart from my time, one way that this manifests is in the generosity of the people who gave freely of their time to assist in the second-marking and other data collection for the evaluation. These people certainly agreed to do so in part, or perhaps wholly, because it was for my PhD thesis, and because of the part that degree plays in my future plans. It would not have been reasonable to ask so much of unpaid volunteers without this additional layer, so that must be acknowledged as part of the design of this experiment. The impact of the findings on my future practice is potentially greater due to the extra evaluation completed because this is for my PhD.

Moore is concerned that findings should lead to “scholarly publication” (p. 5). This thesis hopefully counts, though its distribution will be small. Lists of scholarly talks and publications are given in appendices A and B, and more will be undertaken after the submission of this thesis.

6.8 Concepts and tests

This section will discuss aspects of the statistical analysis that will be used to investigate the questions described above. How these tests will be used in practice will be covered in chapter 7.

6.8.1 Properties of assessment marks

Assessment marks can be considered as being on an interval scale. That is, individual marks are ordered and distance relationships are meaningful, but the zero is an arbitrary placeholder and does not “imply that a student knows absolutely nothing about the materials covered by [an] examination”. So we intend that the difference between pieces of work scoring 60% and 70% is comparable to the difference between pieces of work scoring 50% and 60%, but it is not meaningful to say that a score of 70% by one student means that student knows twice as much about the topic as a student that scores 35% (Lomax, 2007; p. 11). Consequently, tests that are applied to assessment marks must be appropriate for interval data.

Considering how assessment marks arise, there seems to be no reason to expect

that these will follow a normal distribution. Marks are assigned on a scale between 0% and 100%, though the pass mark at 40% and the maximum grade category – first class – at 70% will likely distort this range. There is no reason to expect that results will form a symmetrical distribution.

For example, marks from 44 students for individual assignment 2 will be used in section 8.3. A histogram of these marks is given in figure 6.1. Examining the histogram, there is no indication that the data might follow a normal distribution.

To be sure that we cannot assume normality, a test can be applied. The Shapiro-Wilk test for normality is “relatively powerful when testing for departures from normality” and tests the null hypothesis that a set of data “is from a normal distribution with unknown mean and variance” against the alternative hypothesis that the data are “not from a normal distribution function” (Panik, 2005; p. 631).

Taking the marks from individual assignment 2, the Shapiro-Wilk normality test in R gives a p-value of 0.0068. This gives evidence at the 1% level to reject the null hypothesis that the marks for individual assignment 2 arise from a normal distribution. Consequently, the tests used to analyse the marks from individual assignment 2 must not assume normality of marks.

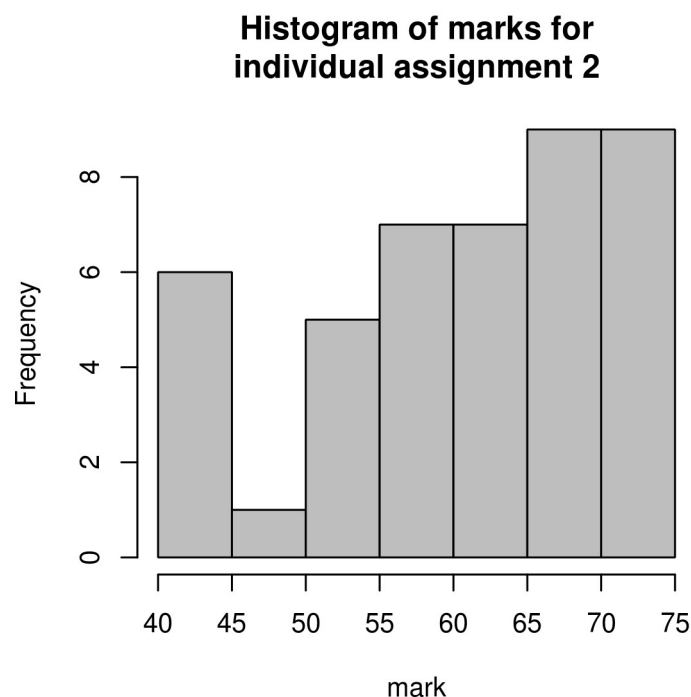


Figure 6.1: Histogram of marks for individual assignment 2.

6.8.2 Marks standardisation

Standardised marks are obtained from raw marks via a process that Bajpai, Calus and Fairley (1978) call coding of data by change of origin and change of unit (p. 36). Say for raw marks $x = (x_1, \dots, x_n)$ we have mean \bar{x} and standard deviation s_x . Then define standardised marks as $X = (X_1, \dots, X_N)$, where

$$X_i = \frac{x_i - \bar{x}}{s_x}.$$

Note that, as a consequence of this definition, $\bar{X} = 0$ and $s_X = 1$.

6.8.3 Measures of correlation

6.8.3.1 Tukey mean-difference plot

The Tukey mean-difference plot takes “a set of (x, y) pairs, and plots the mean $\frac{(x+y)}{2}$ on the x -axis and the difference $x-y$ on the y -axis”. The effect is

“equivalent to rotating the (x, y) clockwise by 45° ” and this is used to “emphasize deviations” from the “positive diagonal” (Sarkar, 2008; p. 208).

So that levels of deviation from perfect correlation can be compared visually between assessments, the Tukey mean-difference plot will be used with standardised marks. Three horizontal lines are also plotted for reference, one dotted line for the mean difference and dashed lines for 1.96 standard deviations above and below the mean difference.

6.8.3.2 Pearson's product-moment correlation coefficient (Pearson's ρ)

Pearson's product-moment correlation coefficient gives a measure of linear correlation between variables. The coefficient, ρ , is between -1 and 1, where zero indicates no linear correlation and a value of 1 or -1 indicates a complete linear relationship (Bajpai, Calus and Fairley; p. 373). It does not assume data are normal (Hyndman, 2010) and is “most appropriate for measurements taken from an interval scale” (Zheng, 2011).

6.8.3.3 Kendall's rank correlation coefficient (Kendall's τ)

Kendall's tau provides a measure of association between ordinal variables, giving a value, τ , between -1 and 1, with zero representing no association and the closer to -1 or 1 representing a stronger association (Cope, 2006; p. 290). The variant tau-b is used where ranked data might include ties (Agresti, 2010; p. 189).

6.8.4 Inter-rater reliability

Comparing exam marks from multiple markers, Massey and Raikes (2006) report the intraclass correlation coefficients (ICC). They use ICC, rather than the Pearson correlation coefficient, because (p. 5)

the intraclass correlation reflects the degree of agreement between two or more examiners, whereas the Pearson correlation reflects the extent to which the relationship between two examiners' marks is linear – a high Pearson correlation would be obtained even if one examiner was consistently more or less severe than the other.

Shrout and Fleiss (1979) define ICC as “the correlation between one measurement (either a single rating or a mean of several ratings) on a target and another measurement obtained on that target” (p. 422). They recommend using intraclass correlation to assess measurement error in judgements made by humans and discuss the selection of an appropriate intraclass correlation coefficient (ICC) model (p. 420).

Shrout and Fleiss describe a “typical reliability study” as: “each of a random sample of n targets is rated independently by k judges” (p. 420). For the experiment here, to measure the extent to which the reliability of marks depends on the actual marker involved, the following applies, that “a random sample of k judges is selected from a larger population, and each judge rates each target, that is, each judge rates n targets altogether” (p. 421). An alternative would consider the reliability of a fixed set of markers, say to assess marker variability using a marking sample ahead of a larger, sub-divided marking task, but here we would like to generalise the result to others within a larger population of markers (pp. 424-5). This situation requires a two-way ANOVA (p. 422) which considers the agreement of multiple markers, in which markers are considered as random effects (p. 425). Shrout and Fleiss's paper is given as a reference for ICC in the 'irr' package in the R manual. Following the advice from Shrout and Fleiss, a two way model on single score data ICC which considers the

agreement between raters will be calculated.

Landis and Koch (1977) give guidance for the interpretation of levels of agreement for kappa statistics. These will be used here for ICC, since Fleiss and Cohen (1973) showed that ICC is a special case of the kappa statistic. The levels proposed by Landis and Koch are (p. 165):

<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

A marker reliability experiment of this sort takes no account of any 'true' value of the mark, if such a thing exists, but only considers the level of agreement between multiple markers. The rating given by ICC is, then, only reliable to the extent that the multiple markers are reliable. For this reason, only people who have marked student work in universities will be used as markers, for a reasonable expectation of reliability.

6.8.5 Comparison of interval or ordinal data from two groups

Where interval or ordinal data has been collected from two independent groups, Fisher's Exact Test can be used to test whether the distribution between categories is independent of group membership, i.e. that the two samples are drawn from equivalent populations (Agresti, 2002; p. 91). If the test gives evidence to reject the null hypothesis, it means that the distribution of responses is significantly different for the two groups. For example, if the same Likert-type question has been asked of two independent groups of students, the null hypothesis is that the distribution of responses between each of the categories (say, Strongly disagree, Disagree, Neutral, Agree and Strongly agree) is independent of which group each student belongs to.

7 Implementation

7.1 Introduction

Recall that students completed a three week group project ('group project 2') based around Art Gallery Problems (see section 5.3), in which they produced a client report and technical report (described in section 5.4). Alongside this, students kept minutes of group meetings and completed two individual assignments, including one that used the partially-automated assessment approach proposed in chapter 4 ('individual assignment 2') (see section 5.5). Following this, students completed a peer assessment of contribution process (see section 5.5).

The partially-automated assessment approach required a technological solution; its implementation for individual assignment 2 will be described here, as will the peer assessment of contribution process. In part, the description of the implementation of the teaching in this module was informed by a reflective log which I kept while delivering it. This approach is recommended by Moore (2011) in order to provide “rich data” (p. 9).

Principles of evaluation were discussed in chapter 6, and their practical implementation will be discussed here. The evaluation aims to answer the main question of whether the proposed partially-automated approach can be put to good educational use in a way that is more advantageous than existing approaches, by answering specific questions about the use of individualised assignments and whether this can impact the issues of plagiarism in coursework and uneven contribution in group work. Eight questions were posed, five of which were particularly in relation to

individual assignment 2:

1. Are the marks particularly sensitive to who is doing the marking?
2. Is the assignment assessing the learning outcomes it was intended to assess?
3. Does the individualised nature of the assignment work to reduce plagiarism?
4. Can the individual work contribute to a view on whether uneven contribution is taking place?
5. Can the individual work contribute to a process for adjusting group marks to take account of individual contribution?

The remaining three were generally about student views of relevant topics:

6. What are student views on conventional measures to address uneven contribution?
7. What are student views on the purpose of assessment?
8. What are student views on automated marking?

Three strands of evaluation were proposed to address these questions.

- Second-marker experiment: to examine reliability (agreement by multiple independent markers) and validity (second opinions on what the assessment was assessing).
- Student feedback questionnaire: to examine student views on uneven contribution and plagiarism in the context of this approach, and on assessment and automated marking generally. Uses both convergent and divergent evaluation approaches.
- Comparison of marks: to examine uneven contribution by comparison of individual assignment marks with group marks and peer assessment of contribution.

7.2 Individual assignment 2

7.2.1 Technology

7.2.1.1 Introduction

The partially-automated assessment approach required a piece of software that would

generate pseudo-randomised question sheets and corresponding answer sheets.

This project, due to its part-time nature, has taken place over an extended time period. One disadvantage of this for a computing project is the rate of change of technology. This has necessitated a shift over time, essentially from a 'home made' approach to an 'off the shelf' one as the project has shifted in emphasis from early plans around technology implementation to later development of research questions around the impact of use of technology.

In many ways, this shift has mirrored changes in the wider mathematics community around technology implementation in assessment. Sangwin (2012) writes that while once “few tools were available to help the CAA system designer”, many more sophisticated software tools and packages have become available “in the last ten years” (i.e. 2002-2012) to assist with development of CAA (p. 3). For example, at the time of initial development it was not reasonable to assume or insist that users would have a JavaScript-enabled web browser (Rowlett, 2004b) or the ability to view MathML (Rowlett, 2005; p. 18). At the time of writing, solutions for viewing MathML via the cross-browser JavaScript system MathJax are very popular (Cervone, 2012; p. 314). MathJax development began in 2009 (Hayes, 2009; p. 101). On the level of systems, the literature reported in section 2.4 detailed the wide variety of bespoke e-assessment implementations, while more recent projects are attempting to draw similarities and promote collaboration with existing systems instead of new development (Barton et al., 2012; p. 20).

Work on this project has broadly followed the pattern of the wider sector, with initial work focused on prototyping technologies that could create a bespoke system, yet the pace of development of underlying technology and mathematical e-assessment systems led ultimately to the adoption of an established system for the automation in this project. This section will briefly outline this process.

7.2.1.2 Design

In section 3.4.5, survey participants were asked about their own use of computers in the assessment process. Half of participants reported using computers to generate examples when setting work, including among Teacher-non-users. Two thirds of all

questionnaire respondents reported using computers to check their answers when setting work, though only about half of Teacher-non-users reported doing so. Just over half of participants said they use computers to check students' answers when marking, though this figure is three quarters for Teacher-users and one quarter for Teacher-non-users. Around three-quarters of participants said they use computers as a marking aid to input and keep track of marks, including among Teacher-non-users.

All this indicates that computers are quite present in the process of setting and marking even traditional, offline assessments for many assessors, and therefore that the use of computers in a system to generate individualised assessments to be taken offline will not be greatly out of place.

The partially-automated approach to assessment in mathematics proposed in chapter 4 required a mathematically-aware system that is capable of producing randomisation in questions and compiling question and answer sheets for printing. Computer interaction would take place in the setting of the assessment template and randomisation options by the lecturer, the output from which would be to be viewed on paper or otherwise offline (perhaps on screen as a PDF) by students for completing the work and the lecturer for marking. Students would submit work without interaction with the system, either offline or essentially offline (perhaps by electronic submission instead of printing). Rather than a full e-assessment system, then, this project required an assessment-setting system.

7.2.1.3 Prototyping

Initial work focused on developing elements for authoring worksheet templates, including input and storage of mathematics notation with dynamically-generated elements, and for translating worksheet templates into individualised, printable worksheets.

Early work on input focused on taking LaTeX code and processing this for storage as MathML, as typing MathML directly is not recommended (Miner, 1998) and LaTeX was already a widely established input format (Hayes; pp. 98-99). Lacking access to a suitably configured or configurable server, attention focused on client-side conversion via LaTeXMathML (see Woodall, 2006), a JavaScript LaTeX to MathML converter

Please check the content of your post.

Test

Some mathematics: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Save

Figure 7.2: The preview stage echoes LaTeX back for LaTeXMathML to interpret.

Test

Some mathematics: $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Published: Thu, 04 Oct 2007 14:01:32 GMT. [permalink](#).

[0 comments](#).

Figure 7.3: The stored MathML code is presented as published content.

I wrote code that would parse the MathML produced by LaTeXMathML and extract the names of any potential constants and variables from questions, essentially by looking for items marked up in MathML using the `<mi>` tag, and present these for the user to set randomisation parameters (see figure 7.4). This worked fairly well, but usability was poor because the system detected any variable and offered it for randomisation. This included names of items the author intended to use as variables and constants, and other items identified by LaTeXMathML using the `<mi>` tag. As well as variable names, Presentation MathML (used by LaTeXMathML) uses the `<mi>` tag for “function names”, including names such as “sin”, “the names of symbolic constants”, such as “ π ”, and “miscellaneous text”, such as “...” (Carlisle et al., 2003).

$$\frac{a}{b-x}$$

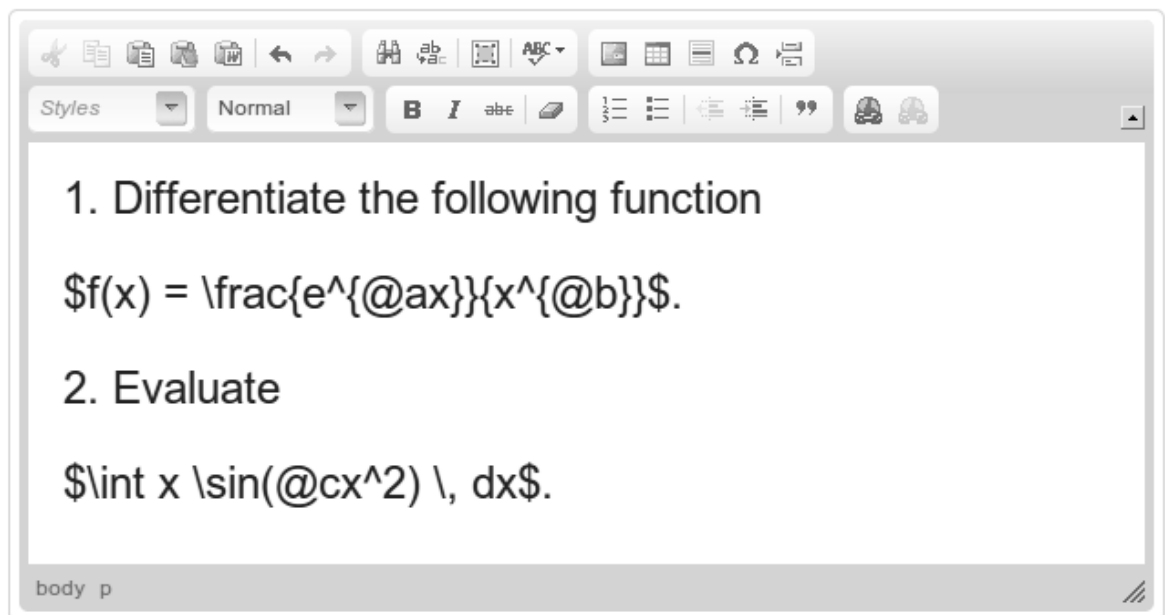
The following can be pseudo-randomised:

a Lower limit: Upper limit:

b Lower limit: Upper limit:

x Lower limit: Upper limit:

Figure 7.4: The variables are echoed back for the user to select which will be pseudo-randomised.



1. Differentiate the following function

$$f(x) = \frac{e^{@ax}}{x^{@b}}.$$

2. Evaluate

$$\int x \sin(@cx^2) \, dx.$$

Figure 7.5: Input via dollar-delimited LaTeX in TinyMCE using @ identifiers for pseudo-randomised constants.

1. Differentiate the following function

$$f(x) = \frac{e^{ax}}{x^b}$$

2. Evaluate

$$\int x \sin(cx^2) \, dx$$

a	Min: <input style="width: 90%;" type="text" value="2"/>	Max: <input style="width: 90%;" type="text" value="9"/>	Dps: <input style="width: 90%;" type="text" value="0"/>
b	Min: <input style="width: 90%;" type="text" value="2"/>	Max: <input style="width: 90%;" type="text" value="9"/>	Dps: <input style="width: 90%;" type="text" value="0"/>
c	Min: <input style="width: 90%;" type="text" value="2"/>	Max: <input style="width: 90%;" type="text" value="9"/>	Dps: <input style="width: 90%;" type="text" value="0"/>

Figure 7.6: Randomisation options for identified constants.

An alternative approach was tried in which the author identifies the constants to be randomised on input by prefixing them with an @ symbol (see figure 7.5). That way, the system knows which of those items using the <mi> tag are to be randomised and which are not. This extra syntax makes writing questions slightly more difficult but the interface is cleaner and overall this method appeared more usable.

Having identified the constants to be pseudo-randomised, these were displayed to the user via a menu interface which asked them to identify both the limits of that randomisation and the number of decimal digits to generate (see figure 7.6). The text and mathematics entered and parameters for pseudo-randomisation formed the content for the worksheet template.

Since the text was in HTML format and the mathematics was in MathML format, it made sense to convert these to XHTML+MathML and embed that in a custom XML format for the worksheet template. This also contained markup noting which parts of the worksheet were questions and answers so that separate question and answer sheets could be generated. This was similar to the approach taken in a previous

project, for which PHP was used to populate the question template with random numbers and compile them into an XHTML+MathML document for online display (Rowlett, 2004a).

Early work to produce printable worksheets focused on conversion from the custom XML worksheet template format via a custom PHP script to RTF format for printing via a word processor. An attempt was also made to produce PDF files by conversion via a custom PHP script to XSL-FO format and processing by the FOP processor (The Apache Software Foundation, 2013) to PDF.

As a trial, this approach was incorporated in a project that aimed to develop individualised worksheets for statistics (Davies and Payne, 2001), including individualised data spreadsheet and elements of statistical analysis produced by the R program (The R Foundation for Statistical Computing, 2013). Davies (2008) describes the worksheets developed by this project as “more interesting and challenging” as well as being used to “encourage individual work” and “to deter plagiarism” (p. 1879).

Though the system worked more or less as intended, there were problems with this approach. The first was the high level of server configuration, particularly installation of software and access rights to run the same from a PHP script. Indeed, following the completion of development the server was replaced by one running a different operating system, which caused much of the code to stop working. A second problem was the limitations of the FOP processor and the consequent limitations on what content could be incorporated into the finished worksheet. Thirdly, the system had user interface issues due to the high level of flexibility given to the worksheet author, who was essentially able to type anything in plain text or R code and mark this up as part of the worksheet (as a question, an answer, code to be run by the system, etc.) (see figure 7.7). Marking up of these areas was done in a rich text area via TinyMCE, but needed to be completed in a way that was XML-compliant. This meant that the user needed to have a good understanding of the technical workings of the system to use it correctly and avoid coding errors. Particularly, it was possible to mistakenly overlap question and answer areas, or produce mismatched nesting of tags in a way that is tolerated by a web browser in HTML but not by FOP in XSL-FO. To avoid these

problems and make the system more resistant would require editing of the TinyMCE code.

The screenshot shows a web-based worksheet template interface. At the top, there are three tabs: 'Worksheet info', 'Data selection', and 'R code'. Below the tabs, there is a form with the following fields: 'Title: Fastfood - Who can do v', 'Topic: Descriptive Statistics', 'Author: Peter Rowlett', and 'Created: 2007-08-07'. Below the form is a rich text editor with a toolbar containing bold, italic, heading, styles, and various icons. The editor contains the following text: 'The data are supplied in CSV format in the CSV file with this worksheet. The worksheet will display 34 columns of data. In this worksheet you will concentrate on analysing the frequency of purchase of fast food (in Column 2), which is a quantitative variable, and the brand last bought (Column 13), which is a qualitative variable. Your analysis will use summary statistics and tables of percentages.' Below the text is a question: 'Q1. Compare the mean and the median number of purchases per month: Do they have different values? If so why do they have different values?'. The question is highlighted with a blue border. Below the question are two sets of answer lines: 'Mean: \$1\$' and 'Median: \$2\$'. The answer lines are highlighted with a red border. At the bottom of the form is a 'Save' button.

Figure 7.7: Worksheet template interface. The question is marked in blue, and the answer in red.

An alternative model of worksheet authoring was attempted. This more restrictive model required the author to specify blocks of content (paragraphs, questions, answers, mathematical expressions, etc.) which were editable one at a time and could be connected together to form a worksheet (see figures 7.8-7.9). This used a JavaScript interface with the script.aculo.us library (script.aculo.us, 2010) for user interface, to edit and rearrange blocks of content. This method appeared to be more robust than the previous system but was awkward to use, particularly when one content type needed to be inserted within another, such as LaTeX within a paragraph.

Finally, a prototype system was developed on a simpler model. This used separate files for the question and answer sheets, and allowed rich text content with mathematics expressed as dollar-delimited LaTeX code and pseudo-randomised constants denoted with an @ symbol. The user would write the question and answer

sheets in a TinyMCE rich text area each (see figures 7.10-7.11) and, following submission, would be invited to choose limits for the indicated pseudo-randomised constants (see figure 7.6). Extra PHP code was written to allow basic calculation with randomised constants by including code within `<? . . . ?>` tags. This was processed by PHP and pre-processed to allow only basic arithmetic.

By this stage in development, use of MathJax was widespread. This meant that the need to process and store MathML code was no longer needed, as the dollar-delimited LaTeX code could reliably be used directly. The system produced templates for question and answer sheets (see figures 7.12-7.13) and used these to generate instances of the question and answer sheets (see figures 7.14-7.15).

By this time, web browsers were much better at printing webpages, which could be formatted using CSS print stylesheets, so that these were now sufficiently well-presented to give to students as a worksheet. Some, particularly Chrome, had incorporated a print option to save a webpage as a PDF document, meaning that the server-side translation could be moved to the client. The use of a CSS page break between worksheets meant that the question sheets in figure 7.14 would be 'printed' one per page, either directly from the web browser or to a PDF document.

The screenshot shows a web interface for adding content blocks. At the top, there are six links: 'Add Text Box', 'Add text input', 'Add LaTeX', 'Add Radio button list', 'Add Stats', and 'Add Checkbox list'. Below these links is a large text input area. The text inside the input area reads: 'A method was sought by which mathematical notation can be included in content created online. A suitable output format for online mathematics is MathML (see Rowlett, 2006). A suitable input format is LaTeX, since a lot of lecturers are familiar with this and the language is fairly intuitive for short snippets of mathematics. The methods outlined here basically convert some content posted through a HTML form to change dollar-delimited LaTeX code, i.e.' Below the text input area is a rich text editor toolbar with buttons for bold (B), italic (I), underline (U), and a list of icons. At the bottom of the interface is a link labeled 'Accept'.

Figure 7.8: Content block input interface showing text input.

Add Text Box	Add text input
Add LaTeX	Add Radio button list
Add Stats	Add Checkbox list

A method was sought by which mathematical notation can be included in content created online. A suitable output format for online mathematics is MathML (see Rowlett, 2006). A suitable input format is LaTeX, since a lot of lecturers are familiar with this and the language is fairly intuitive for short snippets of mathematics. The methods outlined here basically convert some content posted through a HTML form to change dollar-delimited LaTeX code, i.e. $\$...LaTeX\ code...\$, to MathML. What is required is similar to that used in blogging engines so these are explored.$

[Accept](#)

Figure 7.9: Content block input interface showing LaTeX input.

Question sheet

Show help

1. Differentiate the following function

$$f(x) = \frac{e^{ax}}{x^b}.$$

2. Evaluate

$$\int x \sin(cx^2) \, dx.$$

body p

Continue

Figure 7.10: Question sheet input.

Answer sheet
Show help
Show worksheet template
Show parameters

1. $\frac{(@ax-@b)e^{ax}}{x^{<?@b+1?>}}$.$

2. $-\frac{\cos(@cx^2)}{<?@c*2?>}+c$.$

body p

Continue

Figure 7.11: Answer sheet input.

Summary of order

10 copies.

Format:

1. Differentiate the following function

$$f(x) = \frac{e^{ax}}{x^b}.$$

2. Evaluate

$$\int x \sin(cx^2) dx.$$

If you are happy, enter a name to save this set:

Figure 7.12: Interpreted question sheet.

Answer sheets

10 copies.

Format:

$$1. \frac{(ax-b)e^{ax}}{x^{b+1}}.$$

$$2. -\frac{\cos(cx^2)}{c} + c.$$

Figure 7.13: Interpreted answer sheet.

1. Differentiate the following function

$$f(x) = \frac{e^{3x}}{x^9}.$$

2. Evaluate

$$\int x \sin(5x^2) dx.$$

1. Differentiate the following function

$$f(x) = \frac{e^{3x}}{x^4}.$$

2. Evaluate

$$\int x \sin(5x^2) dx.$$

1. Differentiate the following function

$$f(x) = \frac{e^{6x}}{x^4}.$$

Figure 7.14: Generated question sheets.

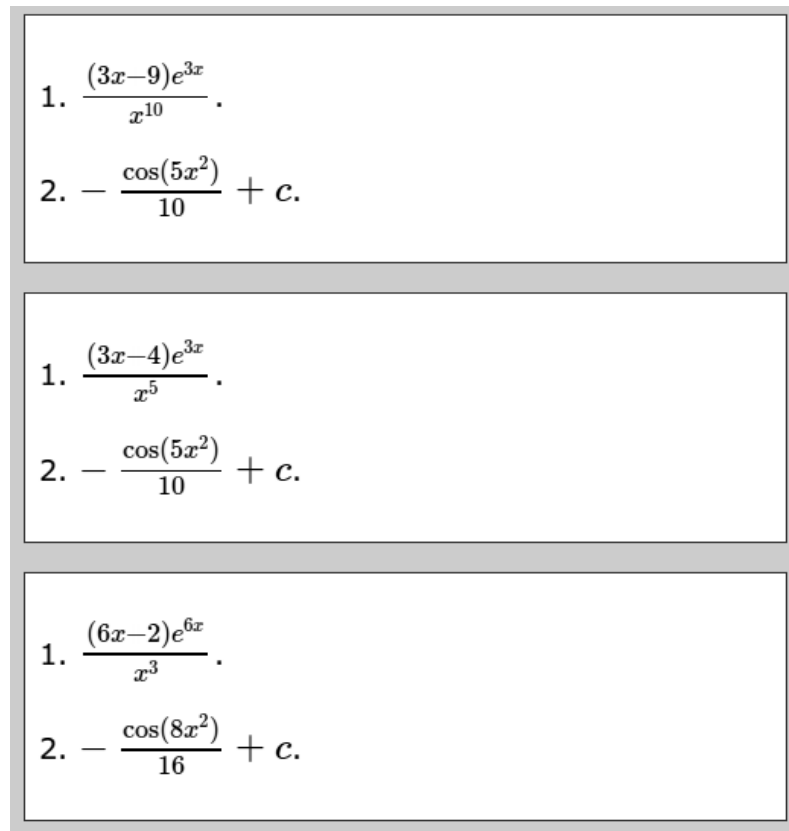


Figure 7.15: Generated answer sheets.

7.2.1.4 Adoption of Numbas

I posted a video on YouTube (see appendix B for details), which gave summary findings from the survey detailed in chapter 3 along with a demonstration of the working prototype shown in figures 7.10-7.15. In response to this, I received the following in an email from Christian Perfect of Newcastle University (Perfect, 2011):

I've written my own e-assessment system, and I think it would do what you want pretty well – there's a print stylesheet which produces minimalistic output similar to what you showed in your video. Have you looked at it? This is exactly the kind of use case my system should be good at, so I'd be willing to do some work to overcome any shortcomings you see.

The system referred to was Numbas, a web-based e-assessment system developed at Newcastle following “the pedagogic design of CALM, CALMAT, SCHOLAR and i-Assess” but “written completely from scratch”. Even though Numbas is principally a fully-featured online e-assessment system, crucially, it was written so that the user interface is “detached from the rest of the system”. This allows “user interface themes” that “can change the way assessments are used”, for example by allowing a “printed worksheet theme” (Perfect, Foster and Youd, 2013).

Advantages of Numbas over my bespoke system include the greater functionality, particularly for randomisation and manipulation of mathematical objects beyond basic arithmetic (Foster, Perfect and Youd, 2012; p. 4), and extensive real-world testing of the system at Newcastle since 2010 (pp. 10-11). This also means that the resultant system is part of a widely-used and actively developed e-assessment system, meaning the potential for future use is much greater than it would be for a bespoke implementation.

Christian Perfect adapted Numbas during 2012 to produce a system which could generate the worksheets needed for individual assignment 2. He provided a statement about the changes made for this project, from which the following quotes are taken. Christian produced

a new theme for Numbas which presents an interface for generating multiple exams at once. Numbas has always been written so the business logic (what I call the 'runtime') is separated from the display code, so this wasn't hard. I didn't have to modify the runtime to accommodate multiple exams existing at once. First draft involved a few stylesheet changes to suit print media, and some buttons to choose how many exams to produce. I already had a print stylesheet for questions, but it would only print one question at a time. The exams are generated and displayed on the screen, then Chrome's print to PDF feature allows you to create PDFs from that.

My feedback on this first iteration led Christian to make further changes, some of which were simply cosmetic. Functional changes were “the ability to print questions and answer sheets separately” and “ID numbers to associate question sheets with answer sheets”. Numbas questions have an 'Advice' section, which is usually used to present feedback to students. Since students will not see the online version, we used the 'Advice' section to input the text for the answer sheet. ID numbers were generated and used as a seed for a deterministic random number generator, `seedrandom.js` (Bau, 2010), so that “putting the same ID in gets you the same randomised worksheet [or answer sheet]”. Christian also added “the ability to create exams in batches, starting from a particular ID number”, with associated controls.

The next section will outline how Numbas was used to generate the worksheets for individual assignment 2.

7.2.2 Implementation of individual assignment 2 in Numbas

In section 5.4, group project 2 was outlined as a project on the topic of Art Gallery Problems, a topic described in section 5.3.3. In section 5.5, individual assignments were proposed in association with the group projects. The module learning outcomes were given in section 5.2.3 as:

1. enhanced problem-solving skills, including the ability to apply mathematical knowledge in real-world scenarios;
2. the ability to work in-depth on a problem over an extended period of time;
3. report writing skills;
4. oral presentation skills;
5. ability to communicate results using different methods;
6. ability to communicate results to audiences of differing mathematical abilities;
7. enhanced team working skills;
8. an appreciation of how groups operate;
9. ability to articulate graduate skills.

In section 5.4, the learning outcomes for group project 2 were given as being numbers 1, 2, 3, 6 and 7. Learning outcome 7 (working in groups) is necessarily only evaluated by the main group project and not the individual work. Learning outcome 6 (communication with different audiences) is assessed in the main project and the reflective essay. Learning outcomes 1 (problem-solving on a real-world scenario), 2 (working in-depth) and 3 (writing reports) are assessed in the main project and in individual assignment 2. Although the depth and amount of report writing are less than in the group project, these skills are still being examined in the individual work. This is because the mathematical assignment requires students to have gained a grasp of the mathematics behind the project via the literature and to have applied this in their individual solutions (learning outcome 2). In writing to relate the theory to the real-world scenario, students are giving verification of some of the skills involved in writing reports (learning outcome 3). By solving the problem and relating this to the real world, students are demonstrating real-world problem solving (learning outcome 1). In conclusion, individual assignment 2 was intended to assess

three learning outcomes:

1. problem-solving skills, including the ability to apply mathematical knowledge in real-world scenarios;
2. the ability to work in-depth on a problem over an extended period of time;
3. some aspects of report writing skills.

To test problem-solving and working in depth, two problems were set: one to solve a set museum and the other to give an example of a floor plan which meets certain criteria. The latter task, asking students to “provide an object satisfying certain mathematical properties”, is regarded by Sangwin (2003) to be a task requiring higher level skills (p. 5). To test skills of relating problem solutions to the real world and report writing, a question asked the students to relate their solution to the real-world scenario. In practice, the question template contained two questions. The first asked students to find the number of guards that would be necessary to guard a museum floor plan and to say whether a certain level of staffing is sufficient to guard the museum 24 hours a day and seven days a week. If so, they must explain how; if not, they must explain why not. The second question asks students to draw a floor plan using a set number of vertices for which a given number of guards is not just sufficient, but also necessary.

The first question used a museum randomly selected from nine floor plans pictured in figure 7.16. Manual sorting was used after printing to ensure that no two students in the same group received the same floor plan. Say the floor plan requires n guards. Part (a) asked for this number of guards. Part (b) asked whether the students could permanently guard the museum pictured with $m = 3n + r$ guards, where r is a random integer equal to 0, 1, 2 or 3. The naïve answer might be that the museum could be guarded with three eight-hour shifts of n guards each, leaving the r guards spare; however, real-world factors make m guards insufficient. Usually, each shift will need one extra guard to allow a rotating break system and more guards to cover leave and illness. If the student answers incorrectly to part (a), their answer will almost certainly be greater than n , so the number in part (b) is still insufficient. If they propose a solution which requires fewer than n guards, i.e. a number which is not even sufficient, a serious misunderstanding has been demonstrated.

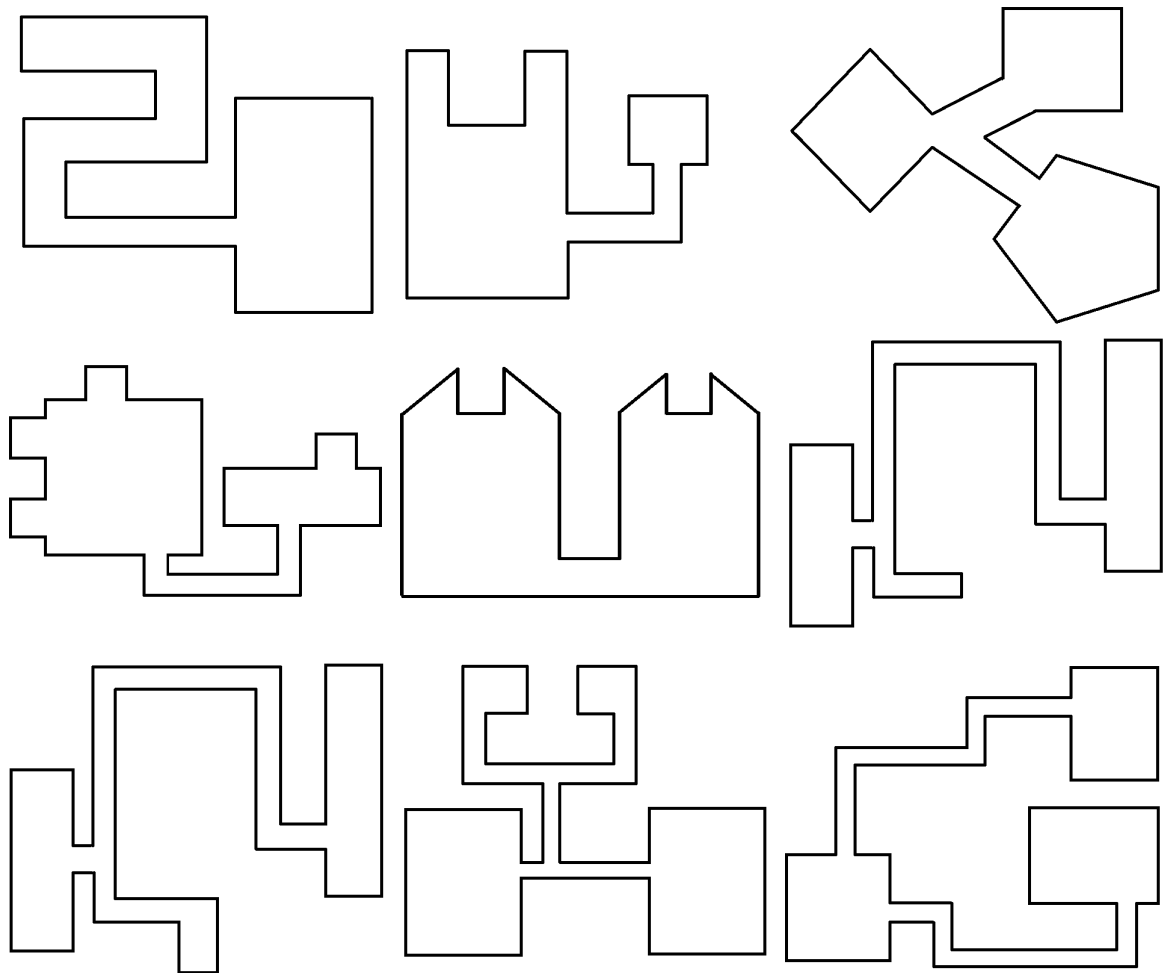


Figure 7.16: Nine museum floor plans, each requiring 3, 4 or 5 guards.

Question 2 asked students to generate a museum with p vertices, where p is a random integer between 10 and 20, satisfying the condition that $\left\lceil \frac{p}{3} \right\rceil$ guards are necessary. In part, this is a test of the depth of their engagement with the topic, since there is an example format widely used in the literature which requires this maximum number of guards per vertex (see figure 5.4).

This means that question 1 (a) effectively used random selection, while questions 1 (b) and 2 used random generation.

A sample piece of randomised work is included in appendix C.

In Numbas, each question is created as a separate object. An 'exam' is then generated which contains both questions. The question statement is written in a rich text interface which would accept dollar-delimited LaTeX (though, in fact, none was used in this case) and pseudo-randomised constants (see figure 7.17 for the example

of question 1; question 2 was similar). The text for the answer sheet is entered in the 'advice' field. Where possible, this was used to give the correct answer; where a correct answer cannot be calculated automatically, a hint of the question that had been asked was shown as a marking aid (see figure 7.18). The pseudo-randomised constants are defined as question 'variables' on a separate screen. For question 1 this defined the number of guards needed, which variant floor plan with that answer would be shown in part (a), the URL where this image could be found and the number of guards to be offered as a staff in part (b) (see figure 7.19). An example generated question sheet is shown in figure 7.20 and the corresponding answer sheet is shown in figure 7.21. The PDF version is shown in appendix C.

I handed the question sheets to the students in groups, to ensure that no two members of the same group got the same floor plan, and recorded the names of which student had taken which ID number. This was time-consuming, but I thought it would be more reliable than asking students to write their ID number on their answer sheet. Students completed the work on paper and handed in a hand-written or typeset report.

NUMBAS
Hi, Peter! Log out

Home Help Exams Questions

agp1

You can edit this question.

Author: Peter Rowlett (prowlett)

Progress: Writing in progress

Test Run
Delete
Make a Copy
Download: standalone .zip, SCORM package, source

General Statement Variables Functions & Rulesets Parts (0) Advice Resources (0)

Exams using this question

Toggle rich text editor

B I U Paragraph

{image(url)}

a) Show, by triangulating and three-colouring the polygon, how many guards are necessary to guard every point in the museum shown above at any one time.

b) In reality, given a staff of {m} guards, could you arrange for every point in the museum shown to be guarded 24 hours a day and seven days a week? If so, how would you arrange this? If not, why not and how many staff would you require?

Path: p

© 2012 School of Mathematics & Statistics, Newcastle University
Some icons from Iconic Icon Set © P.J. Onori, Licensed under Creative Commons Attribution-ShareAlike 3.0 Unported.

Figure 7.17: Numbas question input interface showing question 1.

Toggle rich text editor

B I U Paragraph

a) {n} guards (variant {a}).
b) asking about {m} guards on staff.

Path: p

Figure 7.18: Numbas answer input showing question 1.

agp1

You can edit this question.

Author: Peter Rowlett (prowlett)

Progress: Writing in progress

Test Run

Delete

Make a Copy

Download: standalone .zip, SCORM package, source

General

Statement

Variables

Functions & Rulesets

Parts (0)

Advice

Resources (0)

Exams using this question

+

↑

↓

n

✖ 3

Definition

1 random(3..5)

+

↑

↓

a

✖ 1

Definition

1 random(1..3)

+

↑

↓

m

✖ 9

Definition

1 3*n+random(0..3)

+

↑

↓

url

✖ http://url_hidden/agp/3-1.png

Definition

1 'http://url_hidden/agp/'+n+'-'+a+'.png'

Add another variable

Automatically recalculate variables: ☒

Regenerate values

Figure 7.19: Numbas question 'variable' definition showing variables for question 1.

Generate exam, (starting with ID no.).

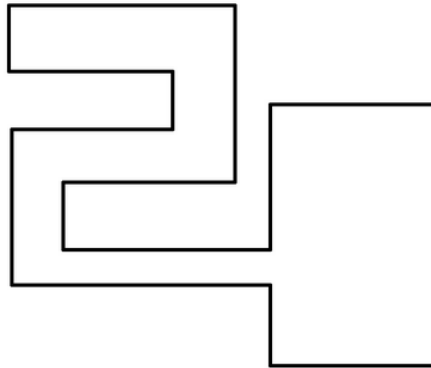
Now

- ☒ Question sheets
- ☐ Answer sheets

ID: 1

MAT30029: Individual assignment 2

1.



- a) Show, by triangulating and three-colouring the polygon, how many guards are necessary to guard every point in the museum shown above at any one time.
- b) In reality, given a staff of 11 guards, could you arrange for every point in the museum shown to be guarded 24 hours a day and seven days a week?
If so, how would you arrange this?
If not, why not and how many staff would you require?

2. Draw a polygon using 16 vertices for which 5 guards are necessary to guard every point at any one time.

Figure 7.20: Numbas-generated question sheet and worksheet generation controls.

ID: 1

MAT30029: Individual assignment 2

1.

a) 3 guards (variant 1).

b) asking about 11 guards on staff.

2.

using 16 vertices for which 5 guards are necessary

Figure 7.21: Numbas-generated answer sheet.

Marking was completed on paper at a computer which provided access to a spreadsheet to identify which student had taken which ID number and into which I recorded the marks, as well as the answer sheet PDF. Marking required me to check the student's ID number from the list and turn the PDF document to the corresponding page, which was a little more time-consuming than it would be if every student had the same questions. However, I do not feel the process was unreasonably extended by the individualisation. Essentially, the answer to question 1 (a) was 3, 4 or 5, depending which diagram the student had been shown. The answer to 1 (b) was “yes” or “no”, depending on the answer to 1 (a), the assumptions made and how well the answer was reasoned. Question 2 would have been hard to mark (checking whether the student had provided a polygon which met particular conditions set in the question) with or without the minor additional requirement to look up what target number had been set for each student.

7.3 Evaluation

The design of the evaluation method was discussed and the statistical methods to be used were outlined in chapter 6. This section details the practicalities of implementation of the evaluation. The results are given in chapter 8.

7.3.1 Second-marker experiment

7.3.1.1 Method

Each second-marking experiment had a piece of work which was marked by an original marker and at least one second-marker. In section 8.1, the marks assigned by each marker are displayed in a table. To assist with interpretation, any comment on the differences in the marks, a Tukey mean-difference plot comparing each second-marker with the original and the intraclass correlation coefficient (ICC) are presented.

The remainder of this section gives the background to the two reference experiments and the second-marking of individual assignment 2.

7.3.1.2 Written examination reference experiment

The work for this reference experiment arose following an open-book test, taken under examination conditions, as part of a basic mathematical methods module for first year mathematics students. The test comprised five well-focused, short problem questions for which 50 marks were available. A 10% sample of all scripts was checked by a moderator, who had access to the original marks, as part of the usual departmental process. The moderator agreed with the marks awarded in all cases.

I marked a sample of ten scripts without reference to the marks assigned by the original marker but using the same mark scheme (blind second-marking). This mark scheme consisted of a set of worked solutions with individual marks indicated for components of answers and for working. The original marker was an experienced assessor employed at the same university as me.

7.3.1.3 Coursework reference experiment

The work for this reference experiment arose from a piece of coursework. The task was to write an 800-1000 word review of a popular book or textbook on mathematics or the history of mathematics, suitable for a mathematics final year undergraduate or graduate audience. A sample of work had previously been approved via a departmental moderation procedure, conducted with reference to the original marks.

I marked a sample of 14 scripts using the marking criteria but without reference to the original marks (blind second-marking). The marking criteria specified those pieces of information that each review should contain, including a summary of the content of the book, as well as some general subjective criteria around the quality of the writing and level of critical understanding. The original marker was working at a different university with a similar entry requirement to my own.

7.3.1.4 Second-marking of individual assignment 2

Three second-marker volunteers were recruited. Each had experience of marking work at university; one as a senior academic, one as a junior academic and one as a PhD student. The students who submitted the work were at a university which required 300 UCAS points for entry. The second-markers' most recent experiences of

marking were at universities that required 260, 300 and 360 UCAS points.

A 10% sample of student work was anonymised (5 pieces from 44 submitted). This was provided along with grade descriptions, a mark scheme and a sample piece of marked work (written to be correct on the non-subjective points of the mark scheme) as a reference piece since the second-markers were not familiar with the topic. The information provided to the second markers (excluding the work to be marked) is included in appendix C. The following instructions were provided to the second markers:

Questions/tasks:

1. Please mark the work. N.B.:
 - Each student has used a separate, individualised question sheet. These are indexed by ID number (top left) and given in setwork.pdf. Each question sheet has a corresponding answer sheet (giving answers where possible, or a reminder of information in the question if not) given in answers.pdf.
 - Please read the marking-notes.pdf and the marking-scheme.pdf, then take a look at the sample-answer.pdf. Please mark the student work contained in student-work.pdf by entering marks into the spreadsheet of marks blank-mark-sheet.xls and return this spreadsheet of marks to me.
2. What was this assessment assessing? (i.e. guess the learning outcomes, or write a general statement of what a student who gets full marks will have demonstrated their ability to do.)
3. Please give any comments about the process of marking individualised work, if you have any.

Thank you!

Marking the sample (task 1) was designed to test reliability. The second-markers were not given an indication of the marks awarded by the first-marker or by each other. There were difficulties in completely specifying the mark scheme, since some of the marks are for subjective measures and interpretation of student writing. This is a large part of what makes it difficult to use automated marking for this piece of work. These subjective elements are similar to those in the coursework reference experiment.

The question about what the assessment was assessing (task 2) was designed to allow evaluation of validity of the assessment. No other information about the coursework or the module was provided to the second-markers. The learning outcomes for this assignment were outlined in section 7.2.2.

Task 3, commenting on the process of marking individualised work, was designed to help identify any problems with the process of marking the work and to attempt to

examine the efficiency of the process.

7.3.2 Student feedback questionnaire

7.3.2.1 Method

A questionnaire looked to explore student views on uneven contribution, plagiarism and collusion, assessment and automated marking.

As well as the cohort taking the module described here, the same questions about individualised assessment (section 7.3.2.3) were given to a group at a different university as a reference group. The lecturer at that university had also used the technique developed for this project via Numbas for an individualised formative in-class question sheet in a final year digital signal processing module. The lecturer hoped to achieve increased engagement, confidence and understanding among students ahead of a summative assessment, and used an individualised assessment to encourage in-class discussion of the questions without collusion. I helped the lecturer with the implementation but had no contact with the students in this group.

For the students taking my graduate skills module, questionnaires were administered online via Google Docs (Google, 2013) during the final session of the module. This was six weeks after group project 2 had been submitted. For the reference group, questionnaires were administered in the same way at the end of the session in which the individualised assessment was used.

The remainder of this section gives the details of the questions that were asked.

7.3.2.2 Student views on conventional measures to address uneven contribution

Students were asked to say whether they took a named role within their group, and what that role was. This is because their role in group hierarchy may impact on their view of group operations.

In order to compare views on the minutes of meetings to the previous experience detailed in section 5.3, students were asked to rate their response to the question 'How helpful did you feel the minutes of meetings were in ensuring team members

completed their assigned tasks?' on a scale from 1 (Very unhelpful or distracting) to 5 (Very helpful).

Students were asked for comments on the peer assessment of contribution, particularly in relation to evenness of contribution, via an open free-text question: "The peer assessment of contribution is designed to make sure that everyone is committed to contributing equally to the project, or to penalise those that don't. I would appreciate any comments you may have on this aspect."

7.3.2.3 Student views on individualised assessment

In order to check whether students had been aware of the nature of this assignment, students were asked 'Were you aware that your question sheet was different to the one given to other students?' In my module, it would be hard for them not to be, because this was made obvious by the process of recording individual ID numbers when handing out the assignments. Still, if they were not aware, it would indicate that they had done the work in complete isolation from other students.

In order to explore their interaction with others regarding the individual work and the issue of plagiarism and collusion, students were asked to indicate their level of agreement with the following four statements on a five-point Likert scale from "1 - Strongly disagree" to "5 - Strongly agree":

- "I disliked having different questions because I wanted to work together with another student on our answers";
- "I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism";
- "I liked having different questions because it meant that no one could copy from me";
- "If we had been set identical questions, members of our group would have copied answers from other students".

Students were asked to respond 'yes' or 'no' to the following two statements, which were accompanied by a reminder that responses were anonymous:

- "While at university, I have copied work from other students";
- "While at university, other students have copied work from me".

In order to explore students' views on the purpose of assessment, students were asked to select up to three answers from the following list in response to the prompt “What are the most important aspects of assessment?”:

- Testing whether I can apply the methods I have been taught;
- Testing whether I have understood the module content;
- Getting a mark back as quickly as possible;
- Ensuring the marking is fair;
- Getting basic feedback quickly;
- Getting detailed feedback;
- Getting the lecturer's view on my work;
- Giving feedback to the lecturer on how I am doing;
- Minimising the marking time for staff;
- Helping me to gain confidence that I understand the topic;
- Helping me to recognise what I can and can't do;
- Helping me to strategically plan my revision;
- Other (please specify).

These statements were drawn from the findings of chapters 2 and 3, and extra suggestions made by the lecturer who was administering the questionnaire to the reference group.

In order to examine views regarding automated marking, students were asked to answer yes or no to the following statement: “I have taken assessment at university which was automatically marked by computer”. Students were then asked to choose a response to “I would rather my work was marked by...” with options “a human marker” and “a computer”. Students were asked to respond in free-text to “Please say why and expand on your answer”.

7.3.3 Comparison of marks

7.3.3.1 Measure of evenness of contribution to group marks

The correlation of raw group project 2 marks and rankings with individual assignment

2 will be presented via Pearson's ρ and Kendall's τ . The dispersion of marks for individual assignment 2 within each group will be examined by examining the range and standard deviation of the individual marks within each group.

7.3.3.2 Correlation between individual and group marks

A reference experiment will establish a context for correlation between similar pieces of work. This uses marks from a basic mathematical techniques module for first year mathematics students. This included two open-book tests, taken under examination conditions. The topics covered by test 1 (basic differentiation and integration, complex numbers, matrices and determinants) are directly used in the topics covered by test 2 (more advanced integration methods, series expansions and linear systems), so it is reasonable to assume that performance on the two tests would be well correlated in most cases. Correlation between the two sets of marks will be examined using Pearson's ρ .

Correlation will be examined in the same way between the means of the individual marks within each group for individual assignment 2 and the group mark for group project 2. The two measures of correlation will be compared.

7.3.3.3 Comparison with peer assessment of contribution

The findings from the previous sections regarding evenness of contribution will be compared with an examination of the other available measure, the peer assessment of contribution.

Peer assessment of contribution followed a model used by Cooper⁶. Say a group contained n students. Then members of this group were given a questionnaire to fill in privately. This asked them to give their name and to assign $10n$ points among the members of their group (including themselves) to represent the contribution to the project made by each group member. An example of the wording used follows:

6 Cooper, B., presented at 'Group Work Working Group meeting', 13th March 2012, University of Bath. Unfortunately the slides are not publicly available and Cooper has no published writing outlining his approach.

Your name:

You have 90 points to distribute between your group to reflect each person's individual contribution.

For example:

Bob Jones	10	<- Bob contributed an average amount
Sarah Smith	12	<- Sarah contributed more than her teammates
Paul Brown	7	<- Paul contributed less than his teammates

Do not show or discuss this form with the others in your group. When you are finished please fold your form and put it in the envelope without showing anyone.

This resulted in n scores per group member. The mean of these scores, rounded to the nearest integer, was used to scale the overall group project mark; students were told that this was the intention before completing the peer assessment of contribution. Taking the mean softened any extreme opinions without completely removing them. Each student included a rating for themselves. A manual check was made to identify any issues with a group member rating themselves higher or lower than the rest of the group. It was not necessary to do so, but an individual's self-rating might have been removed if it was out of step with the group view. Clearly, if everyone in a group assigned an even score (10) to a group member, the overall score for that group member would be 10. For group project 2, the peer assessment scores were in the range 9-12, the mean was 9.95 and the median and mode were both 10.

The investigation of these two measures of evenness of contribution will use techniques from section 6.8 and be a free exploration of the data, rather than following a pre-determined methodology.

8 Results

8.1 Second-marking

8.1.1 Written examination reference experiment

Table 8.1 contains the original marks and those which I (“PR”) assigned during the second-marking. There were five discrepancies of one or two marks (2% or 4% of the total mark). The Tukey mean-difference plot for the standardised total marks is given in figure 8.1. The intraclass correlation coefficient for the two sets of marks is 0.992. These values are regarded by Landis and Koch (1977) as an “almost perfect” level of agreement (p. 165).

Student	Original marker (%)	PR (%)
1	72	72
2	88	86
3	80	80
4	94	94
5	54	54
6	72	72
7	78	76
8	52	48
9	60	58
10	60	62

Table 8.1: Original and second marking for ten written examinations.

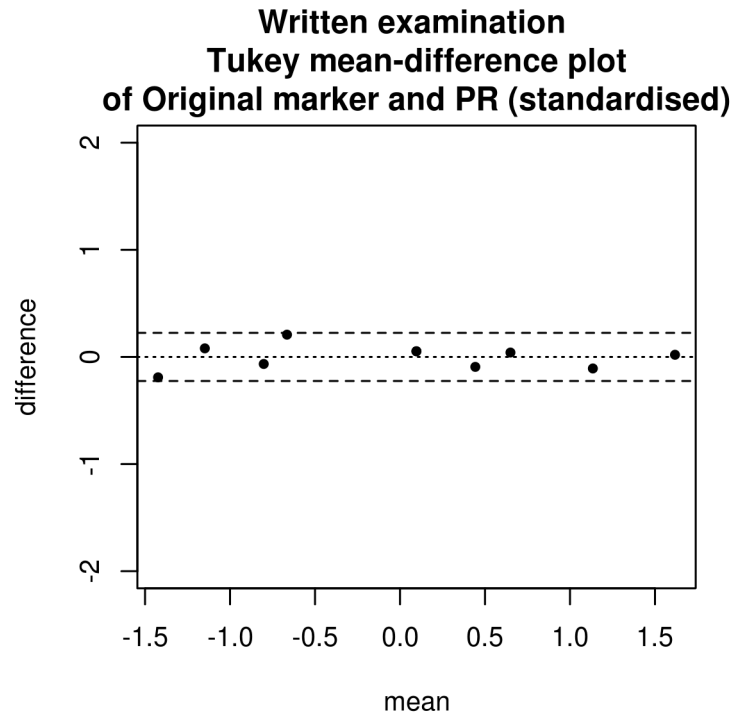


Figure 8.1: Standardised Tukey mean-difference plot for the written examination reference experiment.

8.1.2 Coursework reference experiment

Table 8.2 contains the original marks and those which I (“PR”) assigned during the second-marking. There were differences in all fourteen pieces of work. Six were differences of around 5% or less, a further six were differences of around 10% and two were greater differences. The Tukey mean-difference plot for the standardised total marks is given in figure 8.2. The intraclass correlation coefficient for the two sets of marks is 0.586. This value is regarded by Landis and Koch as a “moderate” level of agreement (p. 165).

Student	Original marker (%)	PR (%)
1	58	56
2	65	55
3	48	43
4	80	71
5	68	45
6	75	74
7	70	58
8	60	63
9	80	68
10	70	58
11	75	61
12	62	58
13	55	51
14	82	72

Table 8.2: Original and second marking for fourteen pieces of open-ended coursework.

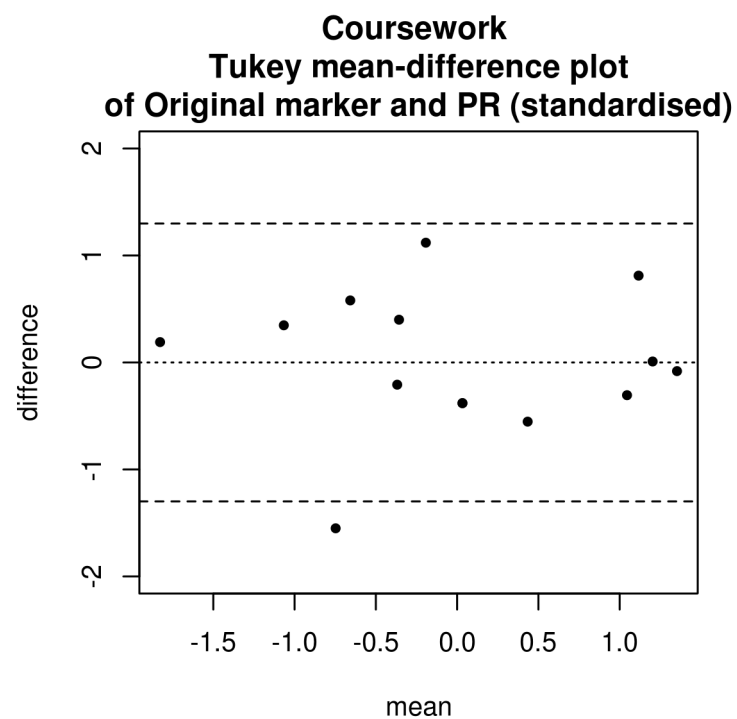


Figure 8.2: Standardised Tukey mean-difference plot for the coursework reference experiment.

8.1.3 Second-marking of individual assignment 2

8.1.3.1 Evaluation of marker consistency

The marks from each marker are given in table 8.3. The Tukey mean-difference plots for the standardised marks are given in figures 8.3, 8.4 and 8.5. The intraclass correlation coefficient for the four sets of marks is 0.635. This value is regarded by Landis and Koch as a “substantial” level of agreement (p. 165).

Student	PR (%)	Second-marker A (%)	Second-marker B (%)	Second-marker C (%)
1	56	31	38	49
2	74	64	59	72
3	67	72	74	77
4	67	46	51	51
5	74	59	54	69

Table 8.3: Original and second marks for five pieces of work submitted for individual assignment 2.

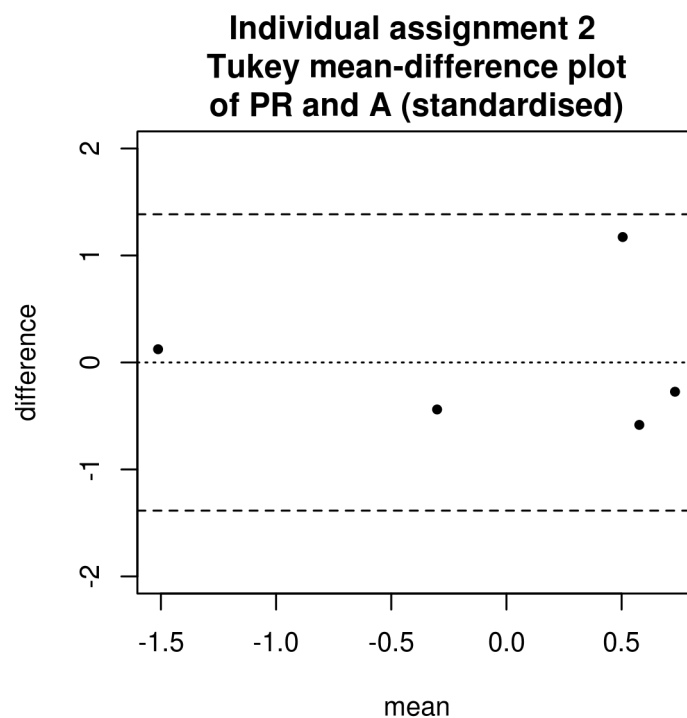


Figure 8.3: Standardised Tukey mean-difference plot with second-marker A for individual assignment 2.

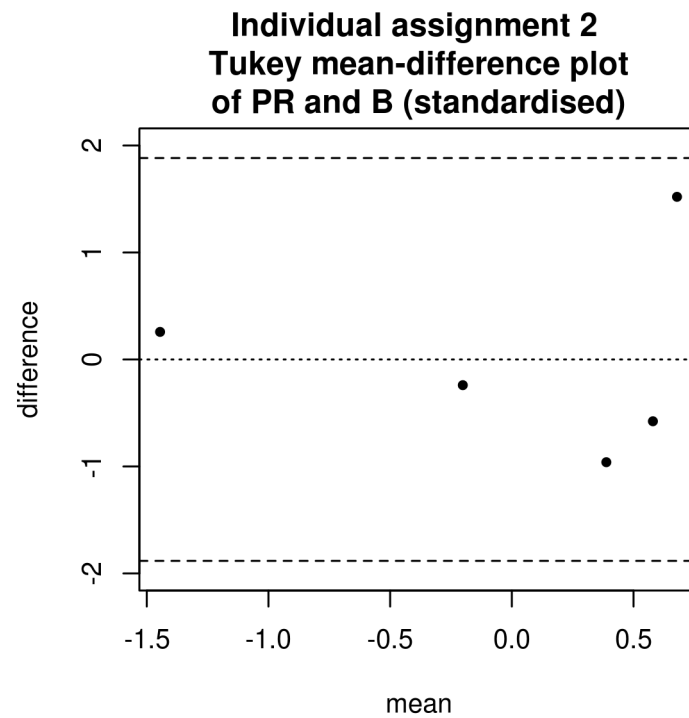


Figure 8.4: Standardised Tukey mean-difference plot with second-marker B for individual assignment 2.

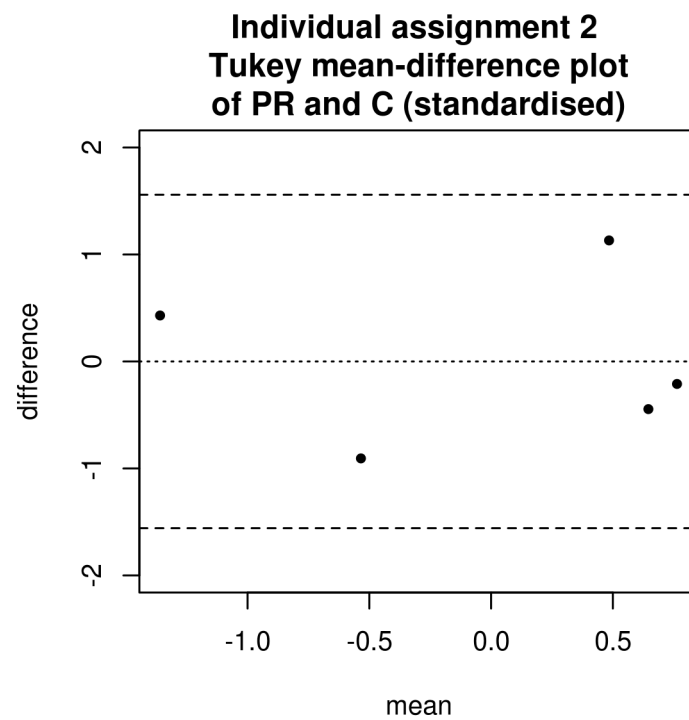


Figure 8.5: Standardised Tukey mean-difference plot with second-marker C for individual assignment 2.

8.1.3.2 Comments on learning outcomes

Second-marker A suggested the following as the learning outcomes that were addressed by this assignment:

- ability to solve unfamiliar problems (or unfamiliar variants of problems discussed in class);
- ability to use literature;
- ability to present mathematical work clearly.

I would suggest that the first of these is problem-solving (learning outcome 1), the second is part of working on a problem in depth (learning outcome 2) and the third relates to communicating results, in this case via a report (learning outcome 3). These three statements do not cover all aspects of the three intended learning outcomes, but neither do they represent additional, unplanned or extraneous requirements.

Marker B gave the following description of what the work was assessing:

The exercise seemed to be designed to assess a student's ability to apply a piece of mathematics, in this case an aspect of computational geometry, and interpret the real world viability of their solution. Particular emphasis was given in the mark scheme for rewarding the students awareness of the mathematical and legal literature as well as communication skills - suggesting that you wanted the students to actually take seriously how one uses mathematics outside of the classroom.

Marker B identified real-world problem-solving and interpreting a mathematical solution outside of the classroom (learning outcome 1), awareness of background information (learning outcome 2) and communication skills (learning outcome 3). Again, these fit with and do not extend the actual intended learning outcomes.

Marker C suggested the following learning outcomes:

- understanding of the theory;
- ability to apply the theory;
- understanding that theory doesn't always apply perfectly to the real world;
- understanding of the difference between 'necessary' and 'sufficient', and that there's more than one possible answer.

Marker C identified aspects of problem-solving, depth of understanding and the process of relating a solution to a real-world scenario (learning outcomes 1 and 2).

Marker C did not identify communication skills (learning outcome 3), nor put forward any additional learning outcomes.

8.1.3.3 Comments on process

Marker A:

I don't think I did this very well. In particular I am unsure about the marks awarded for quality of exposition. To some extent these may have (subconsciously [sic]) overlapped with marks for showing familiarity with literature etc. In general I felt my sample didn't present the work very well. I am a little worried about the marks awarded for consideration of uniqueness and possibility of triangulation, which none of my sample mentioned - was it clear they were expected to address this? Does creating a triangulation show that a triangulation is possible?

I gave a mark to one student in 2 for familiarity with literature although the answer was wrong because there was sufficient similarity with the correct answer to suggest they might have seen it. I was perhaps over-generous. I was probably too harsh elsewhere!

Marker C:

- I had to look up the theory first to understand the question, as I've never come across this problem before
- some of the students seem to miss the point somewhat, and apply the method without really understanding it (i.e. drawing the triangulation but failing to notice by inspection that they have far more guards than necessary)
- none of the students considered that guards are human and might get ill or want to go on holiday!

Marker B gave no comments on the process.

8.2 Student feedback

8.2.1 Student views on conventional measures to address uneven contribution

8.2.1.1 Minutes of meetings

42 students answered the question about minutes of meetings. The responses are shown in table 8.4. Responses are given for all students, as well as for minute takers, who have some responsibility and oversight of this process, and those who took a leadership role. This might be as group leader or as leader of a sub-group, where a group sub-divided the task. Those taking a leadership role have some responsibility for trying to get group members to fulfil the obligations committed to in the minutes.

	All students	Those taking a leadership role	Minute takers	Students not taking a leadership role or acting as minute takers
1 (Very unhelpful or distracting)	0	0	0	0
2	6	1	1	4
3	14	7	0	7
4	14	2	3	9
5 (Very helpful)	8	2	1	5

Table 8.4: Student responses to 'How helpful did you feel the minutes of meetings were in ensuring team members completed their assigned tasks?'

This question was also asked of the students involved in my previous experience of group work, which was evaluated in section 5.3.5. According to Fisher's Exact Test, the data provides no evidence to reject the null hypothesis that the distribution of answers is independent of which group was answering ($p=0.2998$).

8.2.1.2 Peer assessment of contribution

Thirty-six students made comments in response to the prompt about peer assessment of contribution.

Nine of the comments were offering non-specific positive encouragement. A typical example is: "This is a good method because without this if a team member was not contributing enough then normally no one would do anything about it. However this is a discrete [sic] way to do so."

One student commented that "if a member or so does not contribute it should be down the other members of the group to address this themselves." However, the previous experience detailed in section 5.3 and indeed the peer assessment scores indicate that this self-regulation does not always take place.

Twelve of the comments were about the process of awarding scores to peers. The system forced an overall number of points, so that to mark someone as having made an above average contribution require that some others were marked as below average. This was in order to maintain balance and force thoughtful application of the

peer assessment measure. However, some students did not like this. A typical example is: "This is difficult because even though I wanted to reward others at times, it was very difficult to then remove the marks from another member of the team."

Four students complained that the process was sensitive to personal clashes within groups. Three of these had taken a leadership role within the group. The following comment is typical, in that several said that they felt some peer assessment is necessary but worried about sensitivity to personal differences.

A better system is needed. I do not feel this is a fair way of assessing students to give them their final marks. I agree that low contributing individuals should be penalised however, if a dislike by a couple of people towards to one student, even if they are contributing no less than anyone else, could leave them with lower marks.

Another was more direct: "I feel that the peer assessment was marked mainly on how people got on with one another, rather than how much work they did."

Four students commented that different members of the group had contributed unevenly to different areas but, perhaps, evenly overall. One said: "Though some people took a more important role, I felt everyone contributed equally to the workload".

Three students said that it was hard to be objective with your friends. One said "I always feel bad when peer assessing but I have marked down those who don't work as hard so it's a good idea".

Three students said that the existence of peer assessment caused a more even contribution. One student commented: "I like the system and agree that people probably felt committed to contributing due to the possibility of being marked down by their peers". Another said: "Without peer assessment, people would not make much of an effort so others would have to put more effort in and everyone would get the same mark".

Two students complained that others in their group might not be fully aware of the contribution they had made. One said: "because we're in such large groups you don't realise what every member is doing for the majority of the time. So it could seem that one member is not doing a lot when they have done their equal share."

One student commented about the danger of "people marking themselves higher than they should be", but since names were given on the form this was accountable

for, and not, in fact, a problem.

One student said: "I do often feel that people have a tendency to do the 'easy' thing and give everyone equal marks quite often."

8.2.2 Student views on individualised assessment

8.2.2.1 Students in this group project

First, students were asked 'Were you aware that your question sheet was different to the one given to other students?' One student answered "No" and 41 students answered "Yes".

Students were asked to indicate their level of agreement with each of four statements, given with responses in table 8.5. Only one student agreed with the statement "I disliked having different questions because I wanted to work together with another student on our answers," and 28 out of 42 disagreed or strongly disagreed with this statement. 31 out of 42 students agreed or strongly agreed with the statement "I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism"; only one disagreed. 24 out of 42 students agreed or strongly agreed with the statement "I liked having different questions because it meant that no one could copy from me"; 4 disagreed. 24 out of 42 students agreed or strongly agreed that if they had been set identical questions, "members of our group would have copied answers from other students"; 7 disagreed or strongly disagreed.

In response to "While at university, I have copied work from other students", 22 answered "Yes" and 19 answered "No" (one student did not answer). In response to "While at university, other students have copied work from me", 35 answered "Yes" and 7 answered "No".

Responses to the prompt "What are the most important aspects of assessment?" are given in table 8.6 in order of their frequency of choice. Not all students obeyed the requirement to select at most three responses, but all responses are presented in the results here. One student selected "Other" and entered "Giving a realistic view on

strengths and weaknesses”.

1 - “Strongly disagree”	2	3	4	5 - “Strongly agree”
“I disliked having different questions because I wanted to work together with another student on our answers.”				
12	16	13	1	0
“I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism.”				
0	1	10	22	9
“I liked having different questions because it meant that no one could copy from me.”				
0	4	14	17	7
“If we had been set identical questions, members of our group would have copied answers from other students.”				
2	5	11	15	9

Table 8.5: Number of students disagreeing and agreeing for four statements.

Response	Number of students
Testing whether I can apply the methods I have been taught	30
Getting detailed feedback	26
Helping me to recognise what I can and can't do	24
Testing whether I have understood the module content	22
Helping me to gain confidence that I understand the topic	14
Ensuring the marking is fair	13
Getting the lecturer's view on my work	12
Getting basic feedback quickly	7
Getting a mark back as quickly as possible	6
Helping me to strategically plan my revision	6
Giving feedback to the lecturer on how I am doing	4
Other - Giving a realistic view on strengths and weaknesses	1
Minimising the marking time for staff	1

Table 8.6: Student choices for “What are the most important aspects of assessment?”

39 out of 42 students answered “Yes” when asked to respond to the statement “I have taken assessment at university which was automatically marked by computer”, and three answered “No”. In a free-text response, one student said that a (presumably optional) module on partial differential equations running concurrently with this one had used Maple T.A. for the assessment.

35 out of 42 chose “I would rather my work was marked by... a human marker” and seven chose “a computer”. 30 out of 42 students answered the free-text prompt to give a reason. All twelve who did not provide a free-text response had answered “a human marker”.

Of the seven who said they preferred a computer to mark their work, four said this was because they could get instant feedback or marks. One said:

I like that this gives immediate feedback, and that any problems can then be directed to a lecturer. It also means that there is less marking for lecturers, which I am sure they are happy about.

Four students who preferred a human marker also commented positively on the instant feedback offered by automated marking, but felt this was outweighed by other advantages of human marking.

However, one student who preferred a human marker said that once the computer had marked correct work as incorrect, and: “we then had to talk to the lecturer to get our mark changed, so I don't believe he saved himself much time at all”.

Two of the four who preferred a computer marker because this meant instant feedback commented about the need for accuracy in syntax, with one saying:

A computer's good as you get feedback immediately however you usually have to type the answer in perfectly but you should get used to it eventually as it makes you check your answers [sic]

Four students who preferred a human marker complained about syntax and input problems.

The remaining three who said they preferred a computer to mark their work commented on the increased accuracy of automatic marking, referring in turn to variation in marks, accuracy in marks and human error.

Of the 23 students who had selected “a human marker” and provided a free-text response, ten said this was because a human marker provides greater detail in feedback. One said this offers “a personal opinion as opposed to right or wrong”.

Another said: "It takes longer to receive the work back, but for wrong questions you may get an explanation as to why it was wrong".

Six students said that a human marker could be more accurate and more tolerant of small errors. One said "a computer error is more likely than a human error, and a human can take more time to mark and go through an answer thoroughly if needs be". Four students preferred a human marker because they could provide method marks for work that was incorrect.

8.2.2.2 Reference group of students at another university

The lecturer felt that the activity had met the intended goals, saying "I think students' confidence did improve, and they seemed to enjoy the activity". He also said:

Individual questions - even when they are very similar - provide an opportunity for a student to gain confidence that they have understood the method by getting the correct answer. If everyone has the same question they can't be as sure that they've implemented the method correctly - it would be easy to skip a step or part of a step, thinking they've understood it, but managing to get the same answer as everyone else.

In response to 'Were you aware that your question sheet was different to the one given to other students?' all 16 students answered "Yes".

Students were asked to indicate their level of agreement with four statements. Responses are given in table 8.7. Three students of 16 agreed with the statement "I disliked having different questions because I wanted to work together with another student on our answers," and 11 disagreed or strongly disagreed. 12 out of 16 students agreed or strongly agreed with the statement "I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism" and none disagreed. Seven out of 16 students agreed or strongly agreed with the statement "I liked having different questions because it meant that no one could copy from me" and four disagreed or strongly disagreed. Seven out of 16 students agreed or strongly agreed that if they had been set identical questions, "some students would have copied answers from other students", and three disagreed (this question was worded slightly differently to that set for the other group, to take account of the changed circumstances; the other questionnaire referred here to "other members of my group").

1 - "Strongly disagree"	2	3	4	5 - "Strongly agree"
"I disliked having different questions because I wanted to work together with another student on our answers."				
3	8	2	3	0
"I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism."				
0	0	4	6	6
"I liked having different questions because it meant that no one could copy from me."				
2	2	5	3	4
"If we had been set identical questions, some students would have copied answers from other students."				
2	1	6	2	5

Table 8.7: Number of students disagreeing and agreeing for four questions.

In response to "While at university, I have copied work from other students", 5 answered "Yes" and 11 answered "No" (one student did not answer). In response to "While at university, other students have copied work from me", 11 answered "Yes" and 5 answered "No".

Responses to the prompt "What are the most important aspects of assessment?" are given in table 8.8. These are presented in the same order as in the previous section for ease of comparison. No student selected "Other".

11 out of 16 students answered "Yes" in response to the statement "I have taken assessment at university which was automatically marked by computer", and five answered "No".

In response to "I would rather my work was marked by...", 15 out of 16 chose "a human marker" and one chose "a computer".

Thirteen gave a free-text response, including the one student who preferred a computer marker, who wrote "Feedback is quicker, and less prone to interpretation".

Of those who selected that they preferred a human marker, one gave a more complicated free-text response, saying:

Dependent on the style of question. Computer marking is good as responses can be almost instantaneous and gives confidence if you are correct, but cannot provide explanations on why you may have gone wrong. Computer marking is helpful when combined with general feedback, i.e. where the majority of students have problems.

I would always prefer final assessments to be marked by hand, but assessments within modules could be marked by computer to give fast feedback and an indication of which aspects are weak points.

Of the remaining twelve, who all selected that they prefer a human marker, seven said this was because the feedback is more detailed, two referred to the award of method marks and two referred to accuracy of marking.

Response	Number of students
Testing whether I can apply the methods I have been taught	14
Getting detailed feedback	9
Helping me to recognise what I can and can't do	10
Testing whether I have understood the module content	10
Helping me to gain confidence that I understand the topic	10
Ensuring the marking is fair	2
Getting the lecturer's view on my work	6
Getting basic feedback quickly	2
Getting a mark back as quickly as possible	0
Helping me to strategically plan my revision	4
Giving feedback to the lecturer on how I am doing	2
Minimising the marking time for staff	0

Table 8.8: Student choices for "What are the most important aspects of assessment?"

8.2.2.3 Comparison between the two groups

The two groups were asked four Likert-type questions about interaction with other students in tables 8.5 and 8.7. The p-values obtained for each question when comparing the two groups via the Fisher Exact Test are presented in table 8.9. In each case, there is no evidence at the 5% level to reject the null hypothesis that the distribution of answers is independent of the group.

Both groups answered "Yes" or "No" to the statement "While at university, I have copied work from other students". Fisher's Exact Test does not give evidence at the 5% level to reject the independence of the answer distributions and groups ($p=0.1513$).

Question	p-value
"I disliked having different questions because I wanted to work together with another student on our answers."	0.08851
"I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism."	0.6193
"I liked having different questions because it meant that no one could copy from me."	0.1366
"If we had been set identical questions, members of our group would have copied answers from other students."	0.3132

Table 8.9: p-values for Fisher's Exact Test on the four Likert scale questions between the two groups of students.

Both groups answered "Yes" or "No" to the statement "While at university, other students have copied work from me". Fisher's Exact Test does not give evidence at the 5% level to reject the independence of the answer distributions and groups ($p=0.2811$).

Both groups selected the most important items from a list of statements about the purpose of assessment in tables 8.6 and 8.8. Fisher's Exact Test does not give evidence at the 5% level to reject the independence of the answer distributions and groups ($p=0.7906$).

Both groups answered "Yes" or "No" to the statement "I have taken assessment at university which was automatically marked by computer". At the 5% level, Fisher's Exact Test does give evidence to reject the independence of the answer distributions and groups ($p=0.03002$), so we can say there is evidence to suggest the experiences of the two groups on e-assessment are significantly different.

Both groups chose "a human marker" or "by computer" in response to the prompt "I would rather my work was marked by...". Fisher's Exact Test does not give evidence at the 5% level to reject the independence of the answer distributions and groups ($p=0.4231$).

8.3 Marks for individual assignment 2

8.3.1 Measure of evenness of contribution to group marks

The raw group project 2 marks and rankings do not correlate well with the marks and rankings for individual assignment 2 ($\rho=0.230$; $\tau=0.229$).

The range and standard deviation of the individual marks within each group are presented in tables 8.10 and 8.11. Individual marks for each group represent a range of at least 23 marks and up to 31 marks, and have a standard deviation of at least 8.216 and up to 11.411.

Group	Individual assignment 2 marks range for group members
A	31
B	30
C	23
D	28
E	30

Table 8.10: Marks range for individual assignment marks within each group.

Group	Individual assignment 2 standard deviation for group members (3 d.p.)
A	11.411
B	10.706
C	8.216
D	9.584
E	9.513

Table 8.11: Standard deviation for individual assignment marks within each group.

8.3.2 Correlation between individual and group marks

8.3.2.1 Reference experiment

For 74 students, Pearson's ρ is 0.700 (3 d.p.), suggesting a reasonably high level of correlation between the two sets of test marks.

8.3.2.2 Correlation between group mean individual marks and marks for group work

The means of the individual marks within each group are given in table 8.12. Group A had eight members and the other groups each had nine members. The raw group marks for group project 2 (i.e. before scaling according to the peer assessment of contribution) are given in table 8.13. Pearson's ρ between the group mean individual assignment 2 marks and the raw group project 2 marks is 0.734 (3 d.p.). This is comparable to the two sets of test marks compared in the previous section.

Group	Individual assignment 2 marks mean for group members
A	54.25
B	62.89
C	64.67
D	61.11
E	60.67

Table 8.12: Means for individual assignment marks within each group.

Group	Group project 2 raw mark
A	45
B	55
C	75
D	75
E	65

Table 8.13: Raw group marks for group project 2.

8.3.3 Comparison with peer assessment of contribution

Section 8.3.1 indicated that even contribution, according to ability as measured by individual assignment 2, was not present within groups. Section 8.3.2 indicated a reasonable level of correlation between the mean individual assignment 2 marks within each group and the raw group project 2 marks. This suggests that marks for individual assignment 2 might be used as a measure of contribution.

The other measure of evenness of contribution is the peer assessment of contribution. This awarded marks for each students' contribution to the project, though, practically, nearly all of these were either even, one mark above even or one mark below, so we might consider this just as a measure of whether the student made an above average, below average or even contribution.

To compare the peer assessment to a measure derived from the individual assignment would involve comparing the individual marks for each student with the mean mark within their group. If a student has scored well above the mean mark for their group, the individual assignment 2 mark suggests that this student is more capable than average for their group at the learning outcomes examined. We might then infer that such a student would be expected to have made an above average contribution to the quality of the group work.

Quantifying the level of difference in marks from the group average that would be regarded as indicating an above or below average contribution is difficult. Limits must be proposed, since we can reasonably expect virtually no one to score exactly the mean mark for their group (in fact, no one did here).

One approach could be to choose arbitrarily to count marks within one standard deviation of the mean as 'average contribution' and marks outside of this range as either 'above average contribution' or 'below average contribution'. Assessing the marks in this way and comparing with the peer assessment of contribution gives Kendall's τ as 0.013 (3 d.p.). This indicates virtually no correlation at all.

Forcing no even contribution, so that everything above the mean is considered

above average and everything below the mean is considered below average, the correlation is $\tau=0.153$ (3 d.p.). This is considerably better, though it still indicates only weak correlation.

A more systematic approach might be to examine the range of differences between the individual marks and the means, in terms of the standard deviation. In this case, the maximum number of standard deviations above the mean is 1.307 (3 d.p.) and the maximum number of standard deviations below the mean is -1.976 (3 d.p.). Approaching this systematically, we might take marks that are within some proportion of the greater of these as 'even contribution'.

Say for assignment scores $x=(x_1, \dots, x_n)$ we have mean \bar{x} , standard deviation s_x and standardised scores

$$X=(X_1, \dots, X_N); \quad X_i = \frac{x_i - \bar{x}}{s_x}.$$

Then $\bar{X}=0$ and $s_X=1$. Then let the values of X which are considered to represent even contribution be those for which

$$\frac{\max(\max(X), |\min(X)|)}{f} > X_i > -\frac{\max(\max(X), |\min(X)|)}{f}$$

where f determines the fraction to be used. Observe that $f=1$ would mean everything is considered an even contribution and as f increases a value arises, say f_u , beyond which everything is considered uneven.

For individual assignment 2, table 8.14 shows the values of Kendall's τ for different values of f . Figure 8.6 shows a plot of τ against f with f evaluated at steps of 0.1 between 1.1 and 18. Note that $f_u \approx 17.5$ here, i.e. all values of $f \geq 17.5$ yield the same value of τ because at that point no students are considered to have made an even contribution. Here values of f between 7 and 10 yield the closest correlation, but it is not clear that this generalises. All values of f produce weak positive correlation between the two measures of evenness of contribution (the highest value of τ is 0.213).

It is not clear whether the result of this experiment is that the individual assignment scores correlate poorly with actual contribution, or that the peer assessment scores do, or both. It is clear that both cannot correlate well with actual contribution, since both do not correlate well with each other, though it is reasonable

to expect that either measure might.

f	τ (3 d.p.)	f	τ (3 d.p.)
2	0.013	11	0.195
3	0.060	12	0.195
4	0.126	13	0.195
5	0.151	14	0.195
6	0.151	15	0.195
7	0.213	16	0.195
8	0.213	17	0.195
9	0.213	18	0.153
10	0.213		

Table 8.14: Values of Kendall's correlation coefficient for different definitions of the range of even contribution.

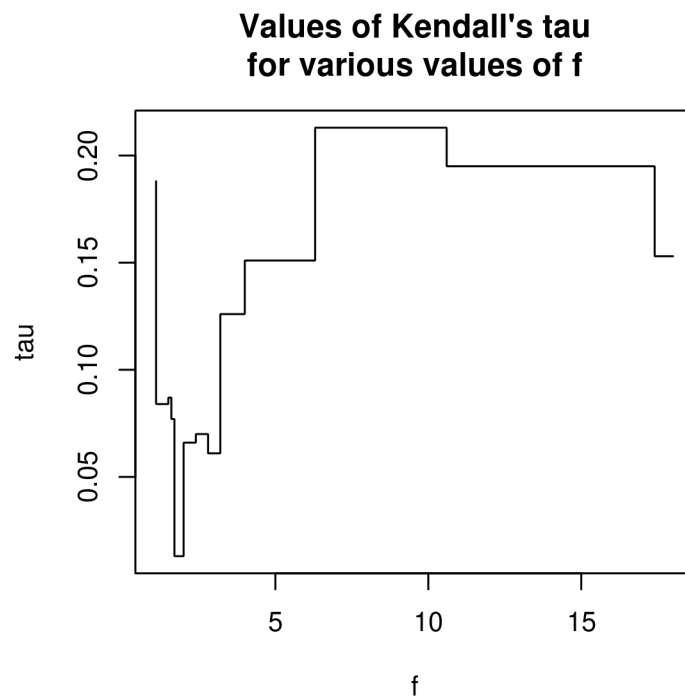


Figure 8.6: Plot of values of Kendall's correlation coefficient for different definitions of the range of even contribution.

9 Conclusions and discussion

9.1 Introduction

Experiences during and alongside this project led to the view that the use of learning technology should be guided by educational need rather than enthusiasm for technology. This research explored assessment and e-assessment in higher education mathematics and proposed a partially-automated approach, novel to higher education mathematics, in which the tools of e-assessment are used to set an individualised assessment that is taken and marked offline. Rather than simply implement this technology and recommend its use, care was taken to identify a teaching and learning context in which this assessment method might be more appropriate than existing approaches. Implementation then focused on the teaching and learning context and evaluation of whether the partially-automated approach met the intended aims of assessment in that context.

This chapter discusses the findings under each of the three research questions, then overall conclusions are drawn. Finally, this chapter discusses limitations of this research and opportunities for future work.

9.2 An examination of e-assessment in higher education mathematics assessment

Research question 1 asked:

1. How is e-assessment used in higher education mathematics, what are its

advantages and limitations, and how does it compare to other assessment methods in common use in higher education mathematics in the UK?

Key sources on assessment in higher education mathematics in the UK were consulted for information on methods in common use. E-assessment was investigated principally through a systematic literature review and a staff survey. Secondly, the opportunity of taking student feedback was used to access student views on this question.

9.2.1 Literature review and survey

Study of key sources on assessment in higher education mathematics led to an evaluation of different assessment methods in terms of the desirable qualities: reliability; validity; fairness and transparency; efficiency; and, usability.

The following methods of assessment were found to be in use in higher education mathematics: written examinations; coursework (short problem coursework assignments which are similar in format to an exam paper and longer, more open-ended coursework assignments or projects); presentations; peer assessment; posters; and, oral assessment. Each of these could be the result of individual or group work.

A systematic literature review of e-assessment was undertaken, which resulted in close study of 97 articles. Understanding the literature to have a bias towards technology enthusiasts, a survey was designed to access the views of those not represented in the literature, including non-users. The literature review and survey were undertaken as a free exploration of e-assessment in higher education mathematics and without a specific goal in mind; they are presented in sections 2.4 and 3 as part of the original contribution of this thesis. A brief synopsis of findings that particularly relate to the later phases of this research is presented below.

E-assessment is most often used for routine work and to test recall, and less often for open-ended or project work, though a computer algebra system can allow free-text entry and access to higher-level skills. Randomisation can allow reuse of questions, including so that different students are presented with different questions in an attempt to dissuade plagiarism. There is a danger of e-assessment leading

students to adopt a surface approach to learning.

Instant feedback is given as a major advantage of e-assessment, particularly for building confidence. However, it is not certain whether a human marker or computer will provide more detailed feedback and neither is it certain which type of marking would be preferred if resources were unlimited. Human markers can be inconsistent, but automated marking cannot mark everything and may be unduly strict with small errors.

Writing questions for e-assessment is difficult, particularly because of the need to second-guess an automated marking system and account for additional learning outcomes imposed by computer input. Lecturers in general do not have the necessary expertise.

There was not a great difference detected between the views found in the literature and those surveyed, though the former are better informed about some of the details of how e-assessment works in practice (see section 3.6 for a comparison).

9.2.2 Student feedback

Two groups were surveyed, including one group with which I had no contact. The only question on which evidence was found of a significant difference between the groups was whether they had been assessed by e-assessment.

Both groups were asked to identify items from a list in response to the prompt “What are the most important aspects of assessment?” The most common responses for both groups were that assessment: tests whether students can apply the methods they have been taught; tests whether they understand the module content; gives detailed feedback to students; and, helps students recognise what they can and cannot do. Interestingly, “getting basic feedback quickly” was not highly rated by students in either group. This is interesting because of the apparent discrepancy between staff and student opinion; instant feedback was cited often as the main advantage of e-assessment in the literature and was highly rated in the staff survey.

In the literature and survey, there was uncertainty over whether a human or computer marker was preferred, and particularly which would give more detailed feedback. Most students surveyed, in both groups, preferred to be marked by a

human marker, even though the survey showed a significant difference between the two groups in experience of e-assessment. The minority of students who preferred to be marked by computer identified instant feedback and marking reliability as advantages. However, user input problems were identified as an issue, and human markers were thought to provide greater detail in feedback and be more forgiving with respect to small errors.

9.3 Is there an opportunity for a novel approach?

Research question 2 asked:

2. Can a novel assessment approach be identified which uses e-assessment techniques to provide a combination of potential advantages and limitations that is distinct from that offered by the commonly used methods identified?

The discussion of assessment methods, in general terms, leads to a view of different methods as each being capable of accessing a different balance of the desirable qualities: reliability; validity; fairness and transparency; efficiency; and, usability. Although the extent to which each is met, and whether this is an issue, will certainly depend on the teaching and learning context in which the method is used, it is fair to say that an open-ended coursework has greater potential for validity, in terms of assessing everything a mathematics degree ought to assess, than a written examination.

A hypothetical written examination might be considered to offer a highly reliable approach. It could, however, be lacking in validity with regard to certain learning outcomes, for example graduate skills development. The assessment could be made more valid if we replace it with a piece of coursework, in which questions could be asked that require greater depth and students could be given more time to answer. The lack of an invigilated environment, however, could lead to concerns over plagiarism, meaning this form of assessment is potentially less reliable. So we bargain increased validity for decreased reliability. The increased validity may make this method more useful for potential employers, but the uncertainty over who completed the work is a problem.

On the other hand, the assessment could be made more reliable than even the written examination, by removing issues of marker inconsistency, if we replace it by an e-assessment. Such a replacement would potentially be even less valid, however, due to the limitations of automated marking, and more difficult for staff to administer.

To be clear, this is all dependent on the teaching and learning context. A written examination might be a perfectly valid approach for a first year module where the student is expected to demonstrate mastery of some set of techniques. Provided the techniques can be marked by computer, the assessment might be equally valid as an e-assessment. Indeed, if the requirement is that students can complete routine calculations in good time, a coursework assignment might even be less valid for that purpose. However, in general the characterisation given above applies.

It is noticeable that setting an e-assessment is made more difficult because randomisation can cause questions to be generated that the automated marking system cannot mark. Beyond simple issues of randomisation taken too far, it is possible to imagine questions that cannot be marked by computer but which could contain randomised elements. This suggests the possibility of computer-set individualised worksheets which are marked manually, a form not found in the literature on assessment reviewed under research question 1.

Such an approach would potentially offer: reduced efficiency of marking but increased validity compared with an examination, by allowing more open-ended questions; increased reliability compared with coursework, by reduction of plagiarism (including via collusion); increased efficiency for setting questions compared with e-assessment, because there is no need to consider what the students can input into a computer or to second-guess the limitations of an automated marking engine. There are, then, advantages and disadvantages to this approach over others, and I would say that no other approach offers quite this arrangement of advantages and limitations.

Given that the approach to be taken demands that learning technology innovation follows educational need, not innovator enthusiasm, the question arises of whether a teaching and learning context exists in which the proposed partially-automated approach is useful, and more advantageous than existing approaches.

9.4 Is there a learning and teaching context in which the proposed novel approach is more effective than existing methods?

Research question 3 asked:

3. Is there a particular teaching and learning context in which the proposed novel approach is more appropriate than existing commonly used methods?

Such a context is found in a final year module intended particularly to develop students' skills of applying mathematics to the real world and communicating their findings. The module includes group project activities based around real-world problem-solving. This uses peer assessment of contribution and individual coursework assignments to attempt to address the issue of uneven contribution among group members. A group project ('group project 2' for this module) asked students to complete a piece of research on Art Gallery Problems and answer three questions using the techniques of that topic. Associated with this was an individual assignment ('individual assignment 2' for this module). As this was intended to assess a subset of the same learning outcomes as the group project, the individual assignment required students to demonstrate an understanding of the topic, solve a fourth question and relate its solution to the real-world context.

The similarity of the individual and group tasks meant an increased risk of in-group plagiarism, since the group members had to solve three problems together and one on their own. Established methods to reduce plagiarism would have been invigilation or randomisation, with the former meaning an exam or class test and the latter meaning e-assessment. However, the deeper, more open-ended form of questions that suited the learning outcomes for the individual assignment would be affected by the limited validity that is possible with either a timed written examination or computer-marked e-assessment. Rather than distort the purpose of the assessment to suit the available methods, the proposed partially-automated approach was used.

The main research question, whether a learning and teaching context exists for the proposed approach, thus became subservient to whether a reliable and valid

assignment could be set in this context, and whether this would be effective in relation to plagiarism and uneven contribution.

Following early experimentation and prototyping, eventual implementation used an adaptation of an existing e-assessment system, Numbas. This had the required functionality, following a small amount of modification by its developer, and using an existing, more functional and well-tested system has advantages over a bespoke development. Numbas was used to set an individualised worksheet, which was printed for students to complete on paper for marking by hand.

Handing out the work was more time-consuming than standard work because of the need to do this in groups, to be sure that the members of each group received each diagram only once, and record which student had taken which ID number. Marking was more time-consuming also, because of the need to check the answers for each student in an answer file, though questions were written to be quite open-ended so this was not a major issue.

Evaluation focused on whether a reliable and valid assessment had been set, to what extent this addressed the issue of plagiarism, and how the individual work adds to the capacity to take account of individual contribution in group work. Evaluation is 'for development' of my teaching practice around using the partially-automated approach in the context of group projects, and 'for knowledge' around whether the partially-automated approach can have an appropriate place in the portfolio of assessment methods in higher education mathematics.

Evaluation took three forms: a second-marker experiment, to test reliability using multiple markers and assess validity; student feedback, to examine student views particularly about uneven contribution, individualised assessment (and assessment and e-assessment in general, discussed above in section 9.2.2); and, comparison of marks, to investigate plagiarism and the contribution of the individual work to an understanding of, and measures to account for, uneven contribution.

9.4.1 Are the marks particularly sensitive to who is doing the marking?

A second-marker experiment was designed to address this question. This had three

volunteers who were asked to blind second-mark a sample of five pieces of student work from individual assignment 2.

Understanding that complete agreement between markers is not possible, particularly for a more open-ended assignment, two reference experiments were also completed in order to provide context for the findings of the main second-marker experiment. Given that level of agreement between markers says nothing about any 'true' value of the marks (if such a value exists), only experienced markers were used.

The first reference experiment examined the level of agreement between two markers using the same mark scheme for an open-book written examination. The two markers showed a high level of agreement (intraclass correlation coefficient, ICC 0.992).

The second reference experiment examined the level of agreement between two markers using the same mark scheme for a reasonably open-ended piece of coursework. The two markers showed a moderate level of agreement (ICC 0.586).

The results of the reference experiments support the idea expressed in the literature and reported in section 2.3 that examination is more reliable than open-ended coursework.

For individual assignment 2, a group of four markers showed a level of agreement that was between the two reference experiments, and close to the open-ended piece of work (ICC 0.635). Given that individual assignment 2 contained subjective and open-ended elements, but was more constrained than the reference piece of coursework, this seems to be in line with expectations and suggests a conclusion that individual assignment 2 was, despite its unusual status as individualised work, not unduly sensitive to who was doing the marking.

It should be noted that one marker reported concern about his reliability and consistency, and another reported difficulty understanding the topic from the model answer. These are issues that could have adversely affected reliability, and having markers who are familiar with the topic and allowing some communication between markers to calibrate expectations would likely have improved reliability.

9.4.2 Is the assignment assessing the learning outcomes it was intended to assess?

The three second-markers of individual assignment 2 were not given information on the learning outcomes of the module or the assignment. They were asked to give their views, based on the assignment brief, mark scheme and student work, of what this was assessing. Two markers identified all three intended learning outcomes at least in part, and the third marker, who was least experienced and in particular only had experience of marking and not of setting work, identified two of the three. None of the markers identified unintended learning outcomes which were being assessed by the assignment. The conclusion, based on this, must be that the assignment was assessing what it was intended to assess, and no more.

9.4.3 Does the individualised nature of the assignment work to reduce plagiarism?

Recall that some sources questioned whether plagiarism as a concern was overblown (Cox, 2011; p. 182). For context, 22 students from 41 confessed that they had copied work from another student at university, and 35 out of 42 said that another student had copied work from them at university⁷. In response to questions about working together, students in this group generally reported not wanting to copy from others and being concerned about others copying from them. They generally appreciated being able to discuss their work with no risk of plagiarism and reported the belief that if identical work had been set then some students would have copied from others. These answers appear to validate my concerns that plagiarism was a high risk in this assignment. The responses from a reference group of students, at another university and with whom I had no contact, are similar (there is no evidence to reject the null hypothesis that they were drawn from similar populations), which suggests that the results are not biased by students' exposure to me as a teacher.

Plagiarism was not detected in the usual marking process.

If members of each group colluded on their work, we would expect to see a

⁷ As the module was optional, this group only represented about one third of the year cohort, so the apparent contradiction here is not necessarily due to response bias.

homogeneity of marks within each group and a relationship between the individual assignment marks and group marks. Since individual assignment 2 marks do not correlate well with group project 2 marks (Pearson's $\rho=0.247$), and the dispersion of marks in each group is high (for each group the range was ~20-30 marks and the standard deviation was ~10), collusion does not appear to have been a big problem.

Since the risk of plagiarism appears to have been high, in this case and in general, and plagiarism has not been detected, we may conclude that plagiarism does happen, but did not happen in this case. This suggests that the individualised nature of the assignment did contribute to a reduction in plagiarism.

9.4.4 Can the individual work contribute to a view on whether uneven contribution is taking place?

Students may view evenness of contribution as referring to effort and whether each member of the group has 'pulled their weight' (MacBean, Graham and Sangwin, 2001; p. 7). This is natural, but at university it is not usual to award marks on the basis of effort alone. Rather, assessment should be aligned to the intended learning outcomes, so that a score on an assessment provides some measure of how well the student has met those learning outcomes. My view of evenness of contribution, therefore, relates to whether each student has individually met the learning outcomes to the same extent. If all contributed evenly, we might expect reasonable correlation between individual assignment 2 and group project 2, since the former assesses a subset of the intended learning outcomes of the latter.

As discussed in the previous section, the range of marks within each group was around 20-30 percentage marks. This is high, relative to, say, the 10 mark gap between most grade boundaries. This suggests that members within each group were not all capable of demonstrating ability at each intended learning outcome to the same extent. This might also indicate that not all group members were capable of contributing equally to the group project learning outcomes.

In addition, the correlation between group mean marks for individual assignment 2 and group marks for group project 2 were compared. For reference, the correlation

was compared with two sets of marks from a group of students taking two class tests on similar topics in the same module. The group mean marks for individual assignment 2 correlate with group marks for group project 2 comparably with the correlation between the marks for the two class tests. However, since the individual assignment 2 marks do not correlate well with those for group project 2, we can conclude that some group members were more capable of achieving the examined learning outcomes (and so were bringing up the mean mark), and uneven contribution was taking place within groups.

9.4.5 Can the individual work contribute to a process for adjusting group marks to take account of individual contribution?

The peer assessment of contribution revealed that students detected uneven contribution in most groups, and the group marks for individuals were adjusted accordingly. In my previous experience of group work I was left feeling quite certain that some groups had carried 'passengers' who had obtained a mark that was not appropriate for their individual ability in relation to the learning outcomes. In this case, marks were adjusted, but a question remains over whether the peer assessment of contribution is the most effective and accurate method to determine the necessary adjustments.

Since the group mean marks for individual assignment 2 are a reasonable predictor of group marks for group project 2, we can assert that in some sense a group average level of ability at the latter has been detected by the former. It might be possible, then, to examine the difference between each individual mark and the group mean mark for individual assignment 2 as a measure of evenness of contribution for each group member. Comparing this measure to that produced by the peer assessment of contribution produces, at best, weak correlation. Since it seems reasonable to expect that ability at meeting the learning outcomes for individual assignment 2 would correlate with ability to meet those same learning outcomes in group project 2, it is difficult to use this information to decide which of the two measures is most accurate. Since it is not possible to know who contributed

what to the work for each group, we can only go so far as to say that there might be potential here.

9.4.6 What are student views on conventional measures to address uneven contribution?

Students in this group project gave similar answers to a question about whether minutes of meetings were a useful tool to the answers given by students in my previous experience. This includes a minority of group leaders not buying into the idea of minutes as a tool for ensuring group members complete their assigned tasks. This effectively validates the position taken that minutes are a useful tool for most but cannot be relied on to ensure even contribution (since peer assessment of contribution demonstrated uneven contribution in most groups despite minutes being in use).

Students were generally positive about the need for some peer assessment of contribution, with the possibility of this both causing more even contribution and detecting when uneven contribution has occurred. There may be some issues with the particular implementation, with some students unhappy at having to balance the contribution of different group members. Concerns were also expressed about the reliability of peer assessment, with objectivity, personal differences and not being aware of the work of each group member being highlighted as issues.

9.5 Final conclusions for this project

A partially-automated approach to the assessment of mathematics in higher education was proposed following a detailed examination of the use of e-assessment in higher education mathematics and its comparison with other assessment methods.

Implementation in a particular teaching and learning context has provided evidence that this approach is capable of setting a valid and reliable assessment. Comparison of the coursework set by the partially-automated approach with other work has provided evidence of similar levels of agreement between multiple markers and decreased risk of plagiarism.

One of the interviewees of Thomlinson, Robinson and Challis (2010a) said that it is “not clear what the real benefit is” of coursework, given that copying is a particular problem among weaker students, and Iannone and Simpson (2012a) report some departments moving away from coursework towards in-class tests. The partially-automated approach appears to be capable of adapting a coursework assignment to make it less sensitive to plagiarism while maintaining its validity, though leading to a reduced efficiency for the marker. By contrast, converting the coursework to a written examination or e-assessment in order to reduce the risk of plagiarism can result in a less valid assessment.

This research provides one context in which the partially-automated approach is suitable and apparently more advantageous than other methods in that context. The partially-automated approach is therefore recommended as an appropriate addition to the repertoire of higher education mathematics assessment methods, particularly in the case where an assessment carries a high risk of plagiarism but issues such as validity make an examination or automated marking system sub-optimal.

In addition to this primary conclusion, several other findings have been drawn during this research.

Some evidence was presented that the views of those writing about e-assessment in the literature may not be as subject to 'enthusiasm bias' as my previous experience suggests and some in the literature report believing (e.g. Sangwin, 2012; p. 12). This conclusion is drawn because views expressed in interviews by those not represented in the literature including non-users of e-assessment, and validated via a wider questionnaire, were broadly in line with the literature findings.

Evidence is presented from two small-scale blind second-marking experiments to support the idea expressed in the literature that a written examination is less sensitive to who is doing the marking than a coursework (e.g. Challis, Houston and Stirling, 2004a; p. 19).

Student feedback was presented which suggested that plagiarism does take place among undergraduate cohorts in UK universities and may therefore be more of a problem than some in the literature report believing (e.g. Cox, 2011; p. 182).

In group work, evidence was presented that students view minutes of group

meetings as a useful tool for ensuring group members complete their assigned tasks, validating the suggestion made by Lowndes and Berry to use minutes (2003; p.21). However, this is not found to be capable of eradicating or accounting for uneven contribution.

Individual work for low stakes on the same learning outcomes as a group project was found to have potential as a tool for examining uneven contribution and possibly accounting for its effects in the group project mark.

The reflection on previous experience described in section 2.2 warned against implementing learning technology out of enthusiasm and with no educational need or evidence of benefit. The literature reviewed in chapter 6 gave one of the goals of educational development research as providing evidence for the evaluator to decide whether to continue with, modify or stop an activity (Thomas, 2009; p. 122). This evaluation has provided sufficient evidence to continue exploring the partially-automated approach in some form. The next section will discuss limitations of this research and opportunities for further work.

9.6 Limitations of this study and opportunities for further work

Assumptions made and numbers responding to surveys and taking part in assessments in the various stages of this study have been reported throughout this thesis. This chapter will focus on limitations that warrant further discussion and areas where follow-on work could take up ideas from this thesis.

9.6.1 The partially-automated approach

If this research had established the conclusion that the partially-automated approach was valid using a bespoke prototype system, then the next steps would be to make that system more robust and increase its functionality through further development, and attempt to make it more widely available for others to use. The fact that the Numbas system is open source, available via Mathcentre (Newcastle University, 2012) and that it now incorporates the offline worksheet functionality means that this approach is already widely available to assessors, greatly improving the potential for

future use by me and others. The survey reported in chapter 3 indicated that computers are used in the process of setting and marking work, which supports the idea that a partially-automated approach would not be out of place for many assessors. Of course, systems other than Numbas could be used to run the same or similar approach.

The ethics of testing live on students (since a laboratory test is insufficient) meant that the well-established peer assessment of contribution was used as the main driver of individual adjustment to group marks, while individual assignment 2 was set at a very low proportion of the module mark. Given that this method has produced a valid and reliable assessment, there is evidence to support increasing this, thus making the overall module mark more responsive to individual ability.

In terms of using the partially-automated approach to reduce plagiarism, it is important to note that individualisation only makes plagiarism more difficult, not impossible. Impersonation will always be an issue without invigilation (which, in practice, means identity checks or an invigilator who knows the students well).

There are other potential uses for the partially-automated approach explored here. Another lecturer also used the Numbas individualised worksheet mode to create worksheets for in-class work (discussed in sections 7.3.2 and 8.2.2.2). The lecturer hoped to achieve increased engagement, confidence and understanding among students ahead of a summative assessment, and used an individualised assessment to encourage in-class discussion of the questions without collusion. For much of mathematics, the hope is that students will take what was delivered in lectures (perhaps a worked example or sample proof) and adapt it to answer the question they have been set. If a student decides to 'copy' work from another student, where the work is individualised, by adapting another students' response to suit their own question, then this may well be perfectly desirable (and, really, a well-written assessment ought to distinguish those who understand the topic from those who are mimicking the worked examples). Thus, individualised work might move students from copying into peer learning, which is potentially a very desirable outcome. The lecturer felt that the activity had met the desired goals. The use of the partially-automated approach for in-class discussion could be explored further.

My experience is that one of the most time-consuming aspects of marking is performing follow-through calculations based on partially incorrect student answers. If the computer was used to set the questions, it should be possible to set up a mathematically-aware marking aid system to allow quick allocation of method marks. This could be used whether the work was individualised or not. This would require further development before it could be attempted.

Other forms of partial-automation may better accommodate the limitations of automated marking. For example, a “semi-automatic approach” was suggested by Sangwin (2012), in which routine calculations were checked by computer before the whole submission was checked by hand. Sangwin says this is a pragmatic approach to get around the limitations of automated marking (p. 15). This is attractive as a way to reduce marking load, but it incurs the costs associated with user input to an e-assessment system without, for the human-marked portions at least, the advantage (for the student) of instant feedback.

9.6.2 E-assessment

Staff, in literature and in the survey, strongly reported instant feedback as an advantage of e-assessment, as it delivers feedback at the point of mistake for improved learning. Students surveyed, however, did not very often rate speed of feedback as being among the most important aspects of assessment (though did rate level of detail in feedback). Further work could explore this issue by asking more detailed questions of a larger sample of students. If found more widely, these differences could be explored by investigating student interactions with e-assessment systems. This is particularly of interest as instant feedback is so much felt to be an advantage for student learning and is often used as a justification for implementing e-assessment. The evidence collected in this project indicates that either staff are wrong in asserting this advantage, or students are not being made fully aware of the advantages they are experiencing.

Similarly, some of the literature reported the view that students would prefer answering questions via computer because of the perception of anonymity, and the literature and survey findings were mixed in views about whether a human or

computer marker would be preferred if resources were unlimited. Again, the students in this study most often preferred a human marker, and this could be explored further. Of course, what students report and what they do in practice may not be the same. It would be interesting to run an experiment where students are individually given a choice whether to submit a piece of work for automated marking and instant feedback or to a human marker.

9.6.3 Student feedback survey

Some of the questions in the student feedback survey might be subject to social desirability response bias (Marsden and Wright, 2010; p. 285). This may be the cause of some options, such as 'helping me to strategically plan my revision', being infrequently selected by students as among the most important aspects of assessment. However, the statements that seem like they would be most susceptible to this bias are "while at university, I have copied work from other students" and "while at university, other students have copied work from me". These prompts produced answers which I expect are different to what the students would report to staff if the survey was not anonymous.

In the question "what are the most important aspects of assessment?", the statement 'testing whether I can apply the methods I have been taught' ranked higher than did 'testing whether I have understood the module content' for both groups of students. Staff might be inclined to take this as evidence of a tendency towards a surface learning approach, but the difference is a subtlety that may not be apparent to all students. Ambiguities such as these could be investigated with additional questions, but here this was not the main focus of the research and the overall burden of the survey was already high. This limits the findings regarding student views of the important aspects of assessment and suggests a topic for further research.

9.6.4 Second-marker experiments

Two pieces of work were used to calibrate expectations about marker agreement. A larger investigation could run many more blind second-marking experiments and

attempt to calibrate a scale for agreement between markers for different methods of assessment, which could be used as a reference in future second-marking-based evaluations. Here, gaining access to more second-marked work was not practical, and the resources to second-mark more work are beyond what is reasonable for a relatively small part of this project.

I have assumed that people who mark work at universities are capable of marking work reliably. One marker commented on his concerns about his own reliability, and another commented about her lack of knowledge about the topic of the assessment. Some markers routinely marked work lower than did others. A larger investigation could have attempted to develop a screening test of reliability for markers before they were involved in the actual second-marking, but here this would have added to the burden on the second-marker volunteers.

Under an actual departmental moderation or second-marking system, markers might discuss their expectations of work that deserves different grades in general terms or even view each other's marks before coming to shared agreement on a mark. Here this was avoided to provide clean blind second-marking. Still, some approach could be imagined where, for example, an independent third party comments to an assessor that their work might be systematically skewed in some way and asks them to reconsider, and this might produce more realistic results. For example, in the case of individual assignment 2, two markers gave a mark below 40% to one piece of work. I could have asked them to re-read the grade descriptions and say whether they really felt the work deserved a failing mark, and ask the other marker (who gave a passing mark) to comment on whether they really felt the work deserved to pass. In this case, I took the first marks returned by each second-marker, because that was the experimental design.

The fact that markers did not necessarily have a good understanding of what quality of work would meet what grades, and that no attempt was made to recruit markers who knew the topic, mean that the discrepancies in marks seen here are likely exaggerated over what would be seen in normal departmental moderation process.

It is interesting to note that during the second-marker experiment, in each case

the first assessor gave marks with a greater mean than did the second-markers. If this result was replicated on a larger scale, exploring the reasons for this would be interesting. For example, in each case, the second-marker knew that the original marks had already been returned to the students. It appears that the marker who interacted with the students and returned the mark was inclined to be more forgiving when marking; though there are other factors, for example the first marker is likely to be more familiar with the topic and general expectations of student ability. If second-markers in general systematically mark more harshly than first-markers, for whatever reason, this could explain and possibly quantify some of the difference in ICC observed between markers.

9.6.5 Uneven contribution

The experiment here found that the individual marks within each group were sufficiently different to conclude that uneven contribution was taking place and plagiarism was not. If, however, the marks in one group had been very similar, then this method would not give evidence to decide whether even contribution or plagiarism had taken place. This is the “grey area” identified by Challis, Houston and Stirling (2004a) in determining at what point discussing “a piece of work they all have to do” has crossed the line into collusion (p. 24).

The issue of what we mean by 'uneven contribution' bears further study. Imagine a situation in which two students, both contributing to a piece of work for which they will share the marks, agree that one of them will solve the mathematical problem and the other will write a report based on the solution. These students might reasonably report that they have each completed an even share of the work. However, if there are two learning outcomes – problem-solving and report writing – then I would say that each student has met one learning outcome only, and the students' view of evenness would not identify this discrepancy. It is possible to imagine a situation, even, where the mathematical analysis is poor but the report is well written. In that case, both students would receive credit for the writing, even though one of them was only responsible for the poor mathematics.

In order to get a clearer picture, students could be asked for more detail in peer

assessment on the contribution of each group member to different aspects of the group project work and different learning outcomes, but there was already a high burden on the students in terms of workload, and this would contribute even more information to be processed fairly into a final mark.

It is difficult or impossible to discover what happened within each group, as methods of observation are either self-reporting or risk distorting what is being observed. For example, lecturer observation of group interactions would certainly affect how students interact, as well as being practically difficult as students continue interactions between sessions, including via restricted-access social networks. Students could be asked to keep reflective diaries, but this forced, ongoing reflection is likely to affect the peer assessment scores, either by making the reflection deeper and so the scores more accurate, or by making the group interactions more sensitive to minor disagreements. As an experiment, students could be shown different ratings of their group contribution profile after the project is complete, say one based on average peer assessment and one based on individual assignment scores, and be asked to say which more closely matches what happened in their group. However, it is not clear that students will be able to make an accurate determination. A piece of work was produced through the collective effort of these students, so there must be some true answer to the question of which students did which parts of the work, but it may not be possible to discover this truth.

Investigation of what criteria students are using to fill out a peer assessment of contribution, and whether this relates to effort or something else, would be interesting.

Peer assessment was criticised in the literature over concerns that students may not be qualified to make the necessary judgements (Iannone and Simpson, 2012a; p. 13) and students in this study raised concerns about whether fellow students could be objective when making judgements. I am not too concerned, therefore, about the lack of strong correlation between peer assessment and individual assignment marks measures. However, this does not automatically mean that individual work on the same topic is any better a measure of evenness of contribution. Note that, strictly speaking, the criticism of peer assessment by students here only really applies to this

particular technique, and there are other methods which could have been used.

The picture is complicated still further because in some ways I want uneven contribution (against different learning outcomes), since a group should play to its strengths and distribute work accordingly. To do so would be good group management. So perhaps the peer assessment should only be used to determine when an uneven amount of effort has taken place. However, awarding extra marks to students who have not met all learning outcomes, simply because they have put in the same amount of effort as others, is not correctly aligning this assessment with its intended learning outcomes. Further study is recommended on group interactions and the relationship between individual ability to demonstrate having met learning outcomes and contribution to group work.

Peer assessment adds additional learning requirements. These may be accounted for in the intended learning outcomes by including provision around reflection, articulation and understanding of how groups operate. It is not clear, though, that peer assessment of contribution is the best method for addressing these learning outcomes. If low-stakes individual assessment on the same learning outcomes could be shown to be an effective method for adjusting group marks to take account of individual contribution, then this would have the advantage that the individual work contributes to the same learning outcomes, rather than imposing additional ones.

If further investigation revealed that individual work was indeed a more suitable measure than peer assessment, or that both could be useful in collaboration, note that this would make the individual work higher stakes. This might encourage students to behave differently in relation to the individual work than when these conclusions were determined, and might further encourage collusion.

9.6.6 Future practice

The literature reviewed in chapter 6 recommended that evaluation of curriculum development and reflection on it should affect future practice and should lead to scholarly publication (Moore, 2011; p. 5).

As well as the publications and talks listed in appendices A and B, I have received acceptance of abstracts for two future conference presentations which are relevant to

this thesis. The first is for a talk at the CETL-MSOR Conference in September 2013 about the design and delivery of the whole graduate skills module. The second is an abstract for a research paper at the British Congress of Mathematics Education (BCME) based on the central idea of this thesis – the partially-automated approach to assessment – and its evaluation. This abstract was subject to peer review to determine whether it merited publication and has been accepted. A full paper is due in September 2013 for presentation and publication at the BCME 8 conference in April 2014.

In terms of future practice, certainly the evaluation has led to a recommendation that there is merit in the partially-automated approach in some circumstances, so it can justifiably be used again. The general question about assessment methods, and the particular implementation in terms of group work, have led to several follow-on questions and potential areas of investigation which can develop into future research. In addition, hopefully this thesis will lead to me qualifying for a PhD, which is a key qualification for university lecturer positions. If so, this project will affect my future ability to practice like nothing else I ever do. Then these findings and suggestions for further work have the potential to influence a lifetime of practice for me, and hopefully others will join me in this.

This is not the end of this story.

References

- Agresti, A., 2002. *Categorical Data Analysis*. Second edition. Hoboken, New Jersey, U.S.A.: John Wiley & Sons.
- Agresti, A., 2010. *Analysis of Ordinal Categorical Data*. Second edition. Hoboken, New Jersey, U.S.A.: John Wiley & Sons.
- Anderson, J.A., 1969. *Real Analysis*. London, U.K.: Logo Press.
- Angus, S.D. and Watson, J., 2009. Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, 40(2), pp. 255-272. DOI: 10.1111/j.1467-8535.2008.00916.x
- Ashton, H.S. and Youngson, M.A., 2004. Creating Questions for Automatic Assessment in Mathematics. *Maths-CAA Series*, February. Available via: www.mathstore.ac.uk/repository/mathscaa_feb2004.pdf [last accessed: 11/07/2013].
- Apache Software Foundation, The, 2013. *The Apache FOP Project*. Available via: xmlgraphics.apache.org/fop/ [last accessed: 11/07/2013].
- Bajpai, A.C., Calus, I.M. and Fairley, J.A., 1978. *Statistical Methods for Engineers and Scientists*. Chichester, U.K.: John Wiley & Sons.
- Baker, B., 2004. Computer-aided learning and assessment of critical path analysis. *MSOR Connections*, 4(3), pp. 32-33. DOI: 10.11120/msor.2004.04030032.
- Barton, S., Foster, B., Levesley, J. and McCabe, M., 2012. Barriers in integrating e-assessment into mathematics teaching and learning for first and second year students. In: D. Waller, ed. *Proceedings of the CETL-MSOR Conference, Coventry University, 5th-6th September 2011*. Birmingham: Maths, Stats and OR Network, pp. 14-20.
- Barton, S. and Rowlett, P., 2011. Using an audience response system - what do the audience DO with the feedback? In: D. Green, ed. *Proceedings of CETL-MSOR Conference 2010, Birmingham, 6th-7th September 2010*. Birmingham, U.K.: Maths, Stats and OR Network, pp. 12-22.

- Bau, D., 2010. Random Seeds, Coded Hints, and Quintillions. *A Dabbler's Weblog*, 30 January. Available via: davidbau.com/archives/2010/01/30/random_seeds_coded_hints_and_quintillions.html [last accessed: 11/07/2013].
- Beevers, C., 2000. Computer Aided Assessment in Mathematics at Heriot-Watt University. *MSOR Connections*, 0(1), pp. 17-19. DOI: 10.11120/msor.2000.00010017.
- Beevers, C., 2006. IT was twenty years ago today... *Maths-CAA Series*, January. Available via: www.mathstore.ac.uk/repository/mathscaa_jan2006.pdf [last accessed: 11/07/2013].
- Beevers, C. and Paterson, J., 2002. Assessment in mathematics. In: P. Kahn and J. Kyle, eds., *Effective Teaching and Learning in Mathematics & its Applications*. London, U.K.: Kogan Page, pp. 49-61.
- Beevers, C.E. and Paterson, J.S., 2003. Automatic assessment of problem-solving skills in mathematics. *Active Learning in Higher Education*, 4(2), pp. 127-144. DOI: 10.1177/1469787403004002002.
- Beevers, C.E., Wild, D.G., McGuire, G.R., Fiddles, D.J. and Youngson, M.A., 1999. Issues of partial credit in mathematical assessment by computer. *ALT-J*, 7(1), pp. 26-32. DOI: 10.1080/0968776990070105.
- Bidgood, P., 2006. Creating Statistical Resources from Real Datasets - The STARS Project. In: A. Rossman and B. Chance, eds. *Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Bahia, Brazil, 2-7 July 2006*. Brazil: International Statistical Institute. Available via: http://www.ime.usp.br/~abe/ICOTS7/Proceedings/PDFs/InvitedPapers/3D3_BIDG.pdf [last accessed: 11/12/2013].
- Blanco, M. and Ginovart, M., 2012. On How Moodle Quizzes Can Contribute to the Formative e-Assessment of First-Year Engineering Students in Mathematics Courses. *Universities and Knowledge Society Journal*, 9(1), pp. 354-370. DOI: 10.7238/rusc.v9i1.1277.

- Bloxham, S., 2009. Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, 34(2), pp. 209-220. DOI: 10.1080/02602930801955978.
- Blyth, B. and Labovic, A., 2009. Using Maple to implement eLearning integrated with computer aided assessment. *International Journal of Mathematical Education in Science and Technology*, 40(7), pp. 975-988. DOI: 10.1080/00207390903226856.
- Broughton, S., Hernandez-Martinez, P. and Robinson, C., 2011. Focus groups to ascertain the presence of formative feedback in CAA. *Proceedings of the British Society for Research into Learning Mathematics*, 31(2). Available via: www.bsrlm.org.uk/IPs/ip31-2/BSRLM-IP-31-2-01.pdf [last accessed: 11/07/2013].
- Brouwer, N., Ekimova, L., Jasinska, M., van Gastel, L. and Virgailaite-Meckauskaite, E., 2009. Enhancing mathematics by online assessments: Two cases of remedial education. *Industry & Higher Education*, 23(4), pp. 277-283. DOI: 10.5367/000000009789346095.
- Brown, G., 2001. *Assessment: A Guide for Lecturers*. York, U.K.: LTSN Generic Centre. Available via: www.heacademy.ac.uk/assets/York/documents/resources/database/id3_Assessment_A_Guide_for_Lecturers.rtf [last accessed: 11/07/2013].
- Carlisle, D., Ion, P., Miner, R. and Poppelier, N., eds., 2003. *Mathematical Markup Language (MathML) Version 2.0*. Second Edition. Available via: www.w3.org/TR/2003/REC-MathML2-20031021/ [last accessed: 11/07/2013].
- Cervone, D., 2012. MathJax: A Platform for Mathematics on the Web. *Notices of the AMS*, 59(2), pp. 312-315. DOI: 10.1090/noti794.
- Challis, N., Gretton, H., Houston, K. and Neill, N., 2002. Developing transferable skills: preparation for employment. In: P. Kahn and J. Kyle, eds., *Effective Teaching and Learning in Mathematics & its Applications*. London, U.K.: Kogan Page, pp. 79-91.
- Challis, N., Houston, K. and Stirling, D., 2004a. *Supporting Good Practice in Assessment*. Birmingham, U.K.: Maths, Stats and OR Network.
- Challis, N., Houston, K. and Stirling, D., 2004b. *Good Practice in Assessment Guide for Students*. Birmingham, U.K.: Maths, Stats and OR Network.

- Challis, N., Robinson, M. and Thomlinson, M., 2010. Introduction: the More Maths Grads Higher Education Curriculum Theme. *In*: M. Robinson, N. Challis and M. Thomlinson, eds. *Maths at University: Reflections on experience, practice and provision*. Birmingham, U.K.: More Maths Grads, pp. 5-6.
- Chelimsky, E., 1997. Thoughts for a New Evaluation Society. *Evaluation*, 3(1), pp. 97-118. DOI: 10.1177/135638909700300107.
- Chirwa, L., 2008. A case study on the impact of automated assessment in engineering mathematics. *Engineering Education*, 3(1), pp. 13-20. DOI: 10.11120/ened.2008.03010013.
- Cook, J., Hornby, J. and Scott, L., 2001. Assessment driven learning. *Maths-CAA Series*, December. Available via: www.mathstore.ac.uk/repository/mathscaa_dec2001.pdf [last accessed: 11/07/2013].
- Cooper, D., 2002. A do-it-yourself approach to Computer-Aided Assessment. *Maths-CAA Series*, August. Available via: www.mathstore.ac.uk/repository/mathscaa_aug2002.pdf [last accessed: 11/07/2013].
- Cope, C., 2006. *Beneath the Surface: The experience of learning about information systems*. Santa Rose, California, U.S.A.: Informing Science Press.
- Corbalan, G., Paas, F. and Cuypers, H., 2010. Computer-based feedback in linear algebra: Effects on transfer performance and motivation. *Computers & Education*, 55(2), pp. 692-703. DOI: 10.1016/j.compedu.2010.03.002.
- Cornish, R., Goldie, C. and Robinson, C., 2006. Computer assisted assessment: How many questions are enough? *Maths-CAA Series*, February. Available via: www.mathstore.ac.uk/repository/mathscaa_feb2006.pdf [last accessed: 11/07/2013].
- Cox, B., 2011. *Teaching Mathematics in Higher Education – the basics and beyond*. Birmingham, U.K.: Maths, Stats and OR Network.
- Croft, A.C., Danson, M., Dawson, B.R. and Ward, J.P., 2001. Experiences of using computer assisted assessment in engineering mathematics. *Computers & Education*, 37, pp. 53-66. DOI: 10.1016/S0360-1315(01)00034-3.

- Crouch, C. H. and Mazur, E., 2001. Peer instruction: ten years of experience and results. *American Journal of Physics*, 69(9), pp. 970-977. DOI: 10.1119/1.1374249.
- Davies, N., 2008. Developments in Promoting the Improvement of Statistical Education. *Bulletin of the ISI*, 62, pp. 1875-1880. Available via: isi.cbs.nl/iamamember/CD7-Lisboa2007/Bulletin-of-the-ISI-Volume-LXII-2007.pdf [last accessed: 11/07/2013].
- Davies, N. and Payne, B., 2001. Web-Created Real Data Worksheets. *MSOR Connections*, 1(4), pp. 15-17. DOI: 10.11120/msor.2001.01040015.
- Davis, L.E., Harrison, M.C., Palipana, A.S. and Ward, J.P., 2005. Assessment-driven learning of mathematics for engineering students. *International Journal of Electrical Engineering Education*, 42(1), pp. 63-72. DOI: 10.7227/IJEEE.42.1.8.
- Delius, G.W., 2004. Conservative Approach to Computerised Marking of Mathematics Assignments. *MSOR Connections*, 4(3), pp. 42-47. DOI: 10.11120/msor.2004.04030042.
- Dreyfus, T., 2002. Advanced Mathematical Thinking Processes. In: D. Tall, ed. *Advanced Mathematical Thinking*. London, U.K.: Kluwer Academic Publishers, pp. 25-41.
- Dufresne, R. J., Gerace, W. J., Leonard, W. J., Mestre, J. P. and Wenk, L., 1996. *Classtalk*: a classroom communication system for active learning. *Journal of Computing in Higher Education*, 7, pp. 3-47. DOI: 10.1007/BF02948592.
- Education Endowment Foundation, 2013. *EEF Welcomes Ben Goldacre Report*. Available via: educationendowmentfoundation.org.uk/news/eef-welcomes-ben-goldacre-report/ [last accessed: 11/07/2013].
- Ekins, J., 2008. Interactive mathematics e-quizzes using *Open Mark*. *MSOR Connections*, 8(3), pp. 21-24. DOI: 10.11120/msor.2008.08030021.
- Entwistle, N. and Ramsden, P., 1983. *Understanding Student Learning*. London, U.K.: Croom Helm.

- Fawcett, L., Foster, B. and Youd, A., 2008. Using computer based assessments in a large statistics service course. *MSOR Connections*, 8(3), pp. 45-48. DOI: 10.11120/msor.2008.08030045.
- Ferrão, M., 2010. E-assessment within the Bologna paradigm: evidence from Portugal. *Assessment & Evaluation in Higher Education*, 35(7), pp. 819-830. DOI: 10.1080/02602930903060990.
- Fleiss, J.L. and Cohen, J., 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient As Measures of Reliability. *Educational and Psychological Measurement*, 33(3), pp. 613-619. DOI: 10.1177/001316447303300309.
- Fletcher, L., Milne, S., Ahmed, S. and Neve, P., 2009. The FETLAR Project – Open Educational Resources and E-Assessment Tools for Mathematics. *4th Irish Workshop on Math Learning and Support Centres, Dublin City University, 4th December 2009*. Dublin, Ireland: Irish Mathematics Learning Support Network. Available via: www.dcu.ie/math/mlc/workshop_nov2009/PaulNeve.ppt [last accessed: 11/07/2013].
- Foster, B., 2007. Using computer based assessment in first year mathematics and statistics degree courses at Newcastle University. *MSOR Connections*, 7(3), pp. 41-45. DOI: 10.11120/msor.2007.07030004.
- Foster, B., Perfect, C. and Youd, A., 2012. A completely client-side approach to e-assessment and e-learning of mathematics and statistics. *International Journal of e-Assessment*, 2(2). Available via: journals.sfu.ca/ijea/index.php/journal/article/viewFile/35/37 [last accessed: 11/07/2013].
- Foster, W.H., 2004. Keeping Control using Computer Aided Assessment. *Maths-CAA Series*, December. Available via: www.mathstore.ac.uk/repository/mathscaa_dec2004.pdf [last accessed: 11/07/2013].
- Fuson, K.C., Kalchman, M. and Bransford, J.D., 2005. Mathematical Understanding: An Introduction. In: M.S. Donovan and J.D. Bransford, eds., *How Students Learn History, Mathematics, and Science in the Classroom*. Washington, D.C., U.S.A.: The National Academies Press, pp. 217-256.

- Goldacre, B., 2013a. *Building Evidence into Education*. London, U.K.: Department for Education. Available via: media.education.gov.uk/assets/files/pdf/b/ben%20goldacre%20paper.pdf [last accessed: 11/07/2013].
- Goldacre, B., 2013b. Teachers! What would evidence based practice look like? *Bad Science*, 15 March. Available via: www.badsience.net/2013/03/heres-my-paper-on-evidence-and-teaching-for-the-education-minister/ [last accessed: 11/07/2013].
- Golden, K., 2003. Coping with Diversity in Recruitment: Can CAA Help? *Maths-CAA Series*, March. Available via: www.mathstore.ac.uk/repository/mathscaa_mar2003.pdf [last accessed: 11/07/2013].
- Goldsmith, P.L., 1975. Thirty statistical quotations. *Journal of Applied Statistics*, 2(1), pp. 52-58. DOI: 10.1080/768371126.
- Google, 2013. *Docs*. Available via: <http://docs.google.com/> [last accessed: 11/07/2013].
- Gray, J., 2007. *ASCIIMathML*: now everyone can type MathML. *MSOR Connections*, 7(3), pp. 26-30. DOI: 10.11120/msor.2007.07030026.
- Green, D.R., Harrison, M.C., Palipana, A.S., Pidcock, D. and Ward, J.P., 2004. The Role of CAA in Helping Engineering Undergraduates Learn Mathematics. *Maths-CAA Series*, November. Available via: www.mathstore.ac.uk/repository/mathscaa_nov2004.pdf [last accessed: 11/07/2013].
- Greenhow, M., 2002. Answer Files --- What more do they reveal? *Maths-CAA Series*, January. Available via: www.mathstore.ac.uk/repository/mathscaa_jan2002.pdf [last accessed: 11/07/2013].
- Greenhow, M., 2008. Mathletics – a suite of computer-assisted assessments. *MSOR Connections*, 8(3), pp. 7-10. DOI: 10.11120/msor.2008.08030007.
- Greenhow, M., 2010. Development of Computer-Aided Assessment of Mathematics for First-Year Economics Students . In: C.M. Marr and M.J. Grove, eds. *Responding to the Mathematics Problem: The Implementation of Institutional Support Mechanisms*. Birmingham, U.K.: Maths, Stats and OR Network, pp. 64-49.

- Greenhow, M. and Gill, M., 2004. Setting objective tests in mathematics using QM Perception. *Proceedings of the 8th Computer Assisted Assessment Conference, Loughborough University, 6th-7th July 2004*. Southampton, U.K.: University of Southampton, pp. 115-126. Available via: caaconference.co.uk/pastConferences/2004/proceedings/Greenhow.pdf [last accessed: 11/07/2013].
- Greenhow, M., Nichols, D. and Gill, M., 2003. Pedagogic issues in setting online questions. *MSOR Connections*, 3(4), pp. 25-28. DOI: 10.11120/msor.2003.03040025.
- Griffin, F., 2004. Designing Randomised Quiz Questions for Mathematics – A Case Study. *Maths-CAA Series*, May 2004. Available via: www.mathstore.ac.uk/repository/mathscaa_may2004.pdf [last accessed: 11/07/2013].
- Gwynllyw, R. and Henderson, K., 2009. DEWIS - a computer aided assessment system for mathematics and statistics. In: D. Green, ed. *Proceedings of the CETL-MSOR Conference, Lancaster University, 8th-9th September 2008*. Birmingham, U.K.: Maths, Stats and OR Network, pp. 38-44.
- Hammond, S., 2010. Report on workshop: e-Assessment In Practice. *MSOR Connections*, 10(1), pp. 52-53. DOI: 10.11120/msor.2010.10010052.
- Hatt, J., 2007. Computer-Aided Assessment and Learning in Decision-Based Mathematics. In: M.B. Nunes and M. McPherson, eds. *Proceedings of the IADIS International Conference on e-Learning, Lisbon, Portugal 6th-8th July 2007*. Lisbon, Portugal: International Association for Development of the Information Society, pp. 382-385.
- Hayes, B., 2009. Writing Math on the Web. *American Scientist*, 97(2), pp. 98-102. DOI: 10.1511/2009.77.98.
- Hermans, D.F.M., 2004. CAA in Context: A Case Study. *Maths-CAA Series*, March 2004. Available via: www.mathstore.ac.uk/repository/mathscaa_mar2004.pdf [last accessed: 11/07/2013].

- Hibberd, S., 2002. Mathematical modelling skills. *In*: P. Kahn and J. Kyle, eds., *Effective Teaching and Learning in Mathematics & its Applications*. London, U.K.: Kogan Page, pp. 158-174.
- Hibberd, S., 2005. Use of Projects in Mathematics. *MSOR Connections*, 5(4), pp. 5-12. DOI: 10.11120/msor.2005.05040005.
- Hunt, N., 2007. Individualized Statistics Coursework Using Spreadsheets. *Teaching Statistics*, 29(2), pp. 38-43. DOI: 10.1111/j.1467-9639.2007.00254.x.
- Hyndman, R., 2010. Re: Pearson's or Spearman's correlation with non-normal data. *Cross Validated*, 19 October. Available via: stats.stackexchange.com/a/3733 [last accessed: 11/07/2013].
- Iannone, P. and Simpson, A., 2012a. A Survey of Current Assessment Practices. *In*: P. Iannone and A. Simpson, eds. *Mapping University Mathematics Assessment Practices*. Norwich, U.K.: University of East Anglia, pp. 3-15.
- Iannone, P. and Simpson, A., eds., 2012b. *Mapping University Mathematics Assessment Practices*. Norwich, U.K.: University of East Anglia.
- Iannone, P. and Simpson, A., 2012c. Performance Assessment in Mathematics: Preliminary Empirical Research. *In*: P. Iannone and A. Simpson, eds. *Mapping University Mathematics Assessment Practices*. Norwich, U.K.: University of East Anglia, pp. 127-135.
- Jones, I.S., 2008. Computer-aided assessment questions in engineering mathematics using MapleTA. *International Journal of Mathematical Education in Science and Technology*, 39(3), pp. 341-356. DOI: 10.1080/00207390701734523.
- Jordan, S., Butcher, P. and Ross, S., 2003. Mathematics Assessment at a Distance. *Maths-CAA Series*, July. Available via: www.mathstore.ac.uk/repository/mathscaa_jul2003.pdf [last accessed: 11/07/2013].
- Judd, K. and Keady, G., 2003. A Tale of One City, and Two CAA Systems. *Maths-CAA Series*, January. Available via: www.mathstore.ac.uk/repository/mathscaa_jan2003.pdf [last accessed: 11/07/2013].

- Kahn, P., 2002. Designing courses with a sense of purpose. In: P. Kahn and J. Kyle, eds. *Effective Teaching and Learning in Mathematics & its Applications*. London, U.K.: Kogan Page, pp. 92-105.
- Kennedy, G.E. and Cutts, Q.I., 2005. The association between students use of an electronic voting system and their learning outcomes. *Journal of Computer Assisted learning*, 21(4), pp. 260-268. DOI: 10.1111/j.1365-2729.2005.00133.x.
- King, S.O. and Robinson, C.L., 2009. 'Pretty Lights' and Maths! Increasing student engagement and enhancing learning through the use of electronic voting systems. *Computers & Education*, 53(1), pp. 189-199. DOI: 10.1016/j.compedu.2009.01.012.
- Klai, S., Kolokolnikov, T. and Van den Bergh, N., 2000. Using Maple and the Web to Grade Mathematics Tests. In: J.C. Kinshuk and T. Okamoto, eds. *Advanced Learning Technology: Design and Development Issues*. Los Alamitos, CA: IEEE Computer Society, pp. 89-92.
- Kotecha, M., 2011. Enhancing students' engagement through effective feedback, assessment and engaging activities. *MSOR Connections*, 11(2), pp. 4-6. DOI: 10.11120/msor.2011.11020004.
- Kounin, J.S., 1970. *Discipline and Group Management in Classrooms*. London: Holt, Rinehart and Winston.
- Landis, J.R. and Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), pp. 159-174. DOI: 10.2307/2529310.
- Lawson, D., 2002. Computer-aided assessment in mathematics: Panacea or propaganda? *CAL-laborate*, 9(1). Available via: ojs-prod.library.usyd.edu.au/index.php/CAL/article/download/6095/6745 [last accessed: 11/07/2013].
- Lawson, D., 2003. An Assessment of i-assess. *MSOR Connections*, 3(3), pp. 46-49. DOI: 10.11120/msor.2003.03030046.
- Levesley, J. and Georgoulis, M., 2010. Computer Aided Assessment and Learning in Mathematics. *Journal for Excellence in Teaching and Learning*, 1. Available via: physics.le.ac.uk/journals/index.php/jetl/article/download/249/119 [last accessed: 11/07/2013].

- Lionheart, B., 2009. *Comparing UK university mathematics departments 2009*. Available via: www.maths.manchester.ac.uk/~bl/maths_league/2009/ [last accessed: 11/07/2013].
- Lomax, R.G., 2007. *An Introduction to Statistical Concepts*. Second edition. Mahwah, New Jersey, U.S.A.: Lawrence Erlbaum.
- Lowndes, V. and Berry, S., 2003. Benefits of using Industrially based Group Projects within Mathematics Programmes. *MSOR Connections*, 3(1), pp. 20-22. DOI: 10.11120/msor.2003.03010020.
- MacBean, J., Graham, T. and Sangwin, C., 2001. *Guidelines for Introducing Groupwork in Undergraduate Mathematics*. Birmingham, U.K.: Maths, Stats and OR Network.
- Manwaring, G. and Calverley, G., 1998. Directing your evaluation. In: J. Harvey, ed. *Evaluation Cookbook*. Edinburgh: Learning Technology Dissemination Initiative, pp. 9-11.
- Marsden, P.V. and Wright, J.D., 2010. *Handbook of Survey Research*. Second edition. Bingley, West Yorkshire, U.K.: Emerald Group Publishing.
- Martin, E. and Greenhow, M., 2004. Setting objective tests in linear algebra using QM Perception. *MSOR Connections*, 4(3), pp. 49-53. DOI: 10.11120/msor.2004.04030049.
- Mason, J., 2002. Reflection in and on practice. In: P. Kahn and J. Kyle, eds., *Effective Teaching and Learning in Mathematics & its Applications*. London, U.K.: Kogan Page, pp. 117-128.
- Massey, A.J. and Raikes, N., 2006. *Item-Level Examiner Agreement*. Cambridge, U.K.: Cambridge Assessment. Available via: www.cambridgeassessment.org.uk/images/111065-item-level-examiner-agreement.pdf [last accessed: 11/07/2013].
- Mavrikis, M. and Maciocia, A., 2003. Incorporating Assessment into an Interactive Learning Environment for Mathematics. *Maths-CAA Series*, June. Available via: www.mathstore.ac.uk/repository/mathscaa_jun2003.pdf [last accessed: 11/07/2013].

- McCabe, M., 2002. LOLA, the Missing Link in Computer Assisted Assessment. *MSOR Connections*, 2(1), p. 7. DOI: 10.11120/msor.2002.02010006.
- McCabe, M., 2009. The exponential growth of mathematics and technology at the University of Portsmouth. *Teaching Mathematics and Its Applications*, 28 (4), pp. 222-227. DOI: 10.1093/teamat/hrp021.
- McCabe, M., Heal, A. and White, A., 2001. Computer Assisted Assessment (CAA) of Proof = Proof of CAA: New Approaches to Computer Assessment for Higher Level Learning. *Maths-CAA Series*, September. Available via: www.mathstore.ac.uk/repository/mathscaa_sep2001.pdf [last accessed: 11/07/2013].
- Miner, R., 1998. Re: Goals verses syntax in MathML. *www-math*, 7 May. Available via: lists.w3.org/Archives/Public/www-math/1998May/0007.html [last accessed 11/07/2013].
- Moore, I., 2011. *Evaluating your Teaching Innovation*. Birmingham, U.K.: National HE STEM Programme.
- Naismith, L. and Sangwin, C., 2004a. Implementation of a Computer Algebra Based Assessment System. *Maths-CAA Series*, October. Available via: www.mathstore.ac.uk/repository/mathscaa_oct2004.pdf [last accessed: 11/07/2013].
- Naismith L. and Sangwin C.J., 2004b. Computer Algebra Based Assessment of Mathematics Online. *Proceedings of the 8th Computer Assisted Assessment Conference, Loughborough University, 6th-7th July 2004*. Southampton, U.K.: University of Southampton, pp. 237-244. Available via: caaconference.co.uk/pastConferences/2004/proceedings/NaismithSangwin.pdf [last accessed: 11/07/2013].
- Newcastle University, 2012. *Numbas*. Available via: www.ncl.ac.uk/maths/numbas/ [last accessed 11/07/2013].
- O'Rourke, J., 1987. *Art gallery theorems and algorithms*. New York, New York, U.S.A.: Oxford University Press, Inc.
- Oliver, M. and Conole, G., 1998. Selecting a methodology. In: J. Harvey, ed. *Evaluation Cookbook*. Edinburgh: Learning Technology Dissemination Initiative, pp. 12-13.

- Panik, M.J., 2005. *Advanced Statistics from an Elementary Point of View*. London, U.K.: Elsevier Academic Press.
- Perfect, C., 2011. *your video, "Advantages and disadvantages of e-assessment"*. 25 September. Personal email quoted with permission.
- Perfect, C., Foster, B. and Youd, A., 2013. *Releasing the power of the web. A completely client-side approach to e-learning and e-assessment in numerate disciplines*. [Lecture to: Transforming Assessment Webinar Series]. 3 April. Available via: www.transformingassessment.com/events_3_april_2013.php [last accessed: 11/07/2013].
- Pidcock, D., Palipana, A.S. and Green, D.R., 2004. The role of CAA in Helping Engineering undergraduates to Learn Mathematics. *Proceedings of the 8th Computer Assisted Assessment Conference, Loughborough University, 6th-7th July 2004*. Southampton, U.K.: University of Southampton, pp. 297-306. Available via: caaconference.co.uk/pastConferences/2004/proceedings/Pidcock_Palipana_Green.pdf [last accessed: 11/07/2013].
- Pinto, J.S., Oliverira, M.P., Anjo, A.B., Vieira Pais, S.I., Isidro, R.O. and Silva, M.H., 2007. TDmat – mathematics diagnosis evaluation test for engineering sciences students. *International Journal of Mathematical Education in Science and Technology*, 38(3), pp. 283-299. DOI: 10.1080/00207390601035476.
- Pitcher, N., 2002. Assessment with Mathwise – Lessons, limitations and learning curves. *Maths-CAA Series*, September. Available via: www.mathstore.ac.uk/repository/mathscaa_sep2002.pdf [last accessed: 11/07/2013].
- Pitcher, N., Goldfinch, J. and Beevers, C., 2002. Aspects of Computer-Based Assessment in Mathematics. *Active Learning in Higher Education*, 3(2), pp. 159-176. DOI: 10.1177/1469787402003002005.
- Pollock, M.J., 2002. Introduction of CAA into a mathematics course for technology students to address a change in curriculum requirements. *International Journal of Technology and Design Education*, 12(3), pp. 249-270. DOI: 10.1023/A:1020229330655.

- Pountney, D., Leinbach, C. and Etchells, T., 2002. The issue of appropriate assessment in the presence of a CAS. *International Journal of Mathematical Education in Science and Technology*, 33(1), pp. 15-36. DOI: 10.1080/00207390110086616.
- QAA [The Quality Assurance Agency for Higher Education], 2006. *QAA guidance on the review of subject benchmark statements*. Gloucester, U.K.
- QAA [The Quality Assurance Agency for Higher Education], 2007. *Subject benchmark statement: Mathematics, statistics and operational research*. Gloucester, U.K.
- Quinney, D., 2010. The Role of E-Assessment in Mathematics. In: P. Bogacki, ed. *Electronic Proceedings of the Twenty-second Annual International Conference on Technology in Collegiate Mathematics, Chicago, Illinois, March 11th-14th 2010*. Norfolk, Virginia, U.S.A.: Old Dominion University, pp. 279-288. Available via: archives.math.utk.edu/ICTCM/VOL22/S093/paper.pdf [last accessed: 11/07/2013].
- Raine, D., 2005. National Teaching Fellowship update: Mathematical Techniques and Problem-Based Learning. *MSOR Connections*, 5(2), p. 14. DOI: 10.11120/msor.2005.05020014.
- Ramsden, P., 2004. Fresh Questions, Free Expressions: METRIC's Web-based Self-test Exercises. *Maths-CAA Series*, June. Available via: www.mathstore.ac.uk/repository/mathscaa_jun2004.pdf [last accessed: 11/07/2013].
- R Foundation for Statistical Computing, The, 2013. *The R Project for Statistical Computing*. Available via: www.r-project.org [last accessed: 11/07/2013].
- Robinson, C.L., Hernandez-Martinez, P. and Broughton, S., 2012. Mathematics Lecturers' Practice and Perception of Computer-Aided Assessment. In: P. Iannone and A. Simpson, eds. *Mapping University Mathematics Assessment Practices*. Norwich, U.K.: University of East Anglia, pp. 105-117.
- Robinson, M., Challis, N. and Thomlinson, M., eds., 2010. *Maths at University: Reflections on experience, practice and provision*. Birmingham, U.K.: More Maths Grads.

- Robinson, P., Salt, D., Milne, S. and Cook, J., 2005. A diagnostic test with an automatically linked revision plan. *Maths-CAA Series*, May. Available via: www.mathstore.ac.uk/repository/mathscaa_may2005.pdf [last accessed: 11/07/2013].
- Rowlett, P.J., 2004a. *Asking Questions With MathML: dynamic treatment of XML and pseudo-randomised mathematics assessment*. MSc dissertation, Nottingham Trent University, U.K.
- Rowlett, P.J., 2004b. Pseudo-randomised CAA by "preprocessing" MathML. *Maths-CAA Series*, September. Available via: www.mathstore.ac.uk/repository/mathscaa_sep2004.pdf [last accessed: 11/07/2013].
- Rowlett, P.J., 2005. Mathematics Teaching & Learning and XML. *MSOR Connections*, 5(1), pp. 35-38. DOI: 10.11120/msor.2005.05010019.
- Samuels, P., 2007. Mathematics Support and New Technologies. *MSOR Connections*, 7(1), pp. 10-13. DOI: 10.11120/msor.2007.07010010.
- Sangwin, C., 2003. New opportunities for encouraging higher level mathematical learning by creative use of emerging computer aided assessment. *International Journal of Mathematical Education in Science and Technology*, 34(6), pp. 813-829. DOI: 10.1080/00207390310001595474.
- Sangwin, C., 2004. Assessing mathematics automatically using computer algebra and the internet. *Teaching Mathematics and its Applications*, 23(1), pp. 1-14. DOI: 10.1093/teamat/23.1.1.
- Sangwin, C.J., 2007. Assessing elementary algebra with STACK. *International Journal of Mathematical Education in Science and Technology*, 38(8), pp. 987-1002. DOI: 10.1080/00207390601002906.
- Sangwin, C., 2012. Computer Aided Assessment of Mathematics Using STACK. *Proceedings of 12th International Congress on Mathematical Education, 8th-15th July, 2012, COEX, Seoul, Korea*. Gangnae-myeon, South Korea: Korea National University of Education. Available via: www.icme12.org/upload/submission/1886_F.pdf [last accessed: 11/07/2013].

- Sarkar, D., 2008. *Lattice: Multivariate Data Visualization with R*. New York, New York, U.S.A: Springer Science+Business Media.
- Schofield, D. and Ashton, H., 2005. Effective reporting for online assessment --- shedding light on student behaviour. *Maths-CAA Series*, February. Available via: www.mathstore.ac.uk/repository/mathscaa_feb2005.pdf [last accessed: 11/07/2013].
- script.aculo.us, 2010. *script.aculo.us*. Available via: script.aculo.us [last accessed: 11/07/2013].
- Shaw, R., 1998. Why Evaluate? In: J. Harvey, ed. *Evaluation Cookbook*. Edinburgh: Learning Technology Dissemination Initiative, p. 8.
- Shrout, P.E. and Fleiss, J.L., 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2), pp. 420-428. DOI: 10.1037/0033-2909.86.2.420.
- Simpson, V. and Oliver, M., 2007. Electronic voting systems for lectures then and now: A comparison of research and practice. *Australasian Journal of Educational Technology*, 23(2). Available via: www.ascilite.org.au/ajet/ajet23/simpson.html [last accessed: 11/07/2013].
- Sims Williams, J. and Barry, M., 2003. TAL - A National Database of Questions - Classification is the Key. *Maths-CAA Series*, May. Available via: www.mathstore.ac.uk/repository/mathscaa_may2003.pdf [last accessed: 11/07/2013].
- Spocke, 2007. TinyMCE 2.1.1.1 Released. *TinyMCE Forum*, 14 May. Available via: www.tinymce.com/forum/viewtopic.php?pid=22257 [last accessed: 11/07/2013].
- Strickland, N., 2002. Alice Interactive Mathematics. *MSOR Connections*, 2(1), pp. 27-30. DOI: 10.11120/msor.2002.02010027.
- Strickland, P., 2001. How sould [sic] a perfect computer aided assessment package in mathematics behave? *Maths-CAA Series*, August. Available via: www.mathstore.ac.uk/repository/mathscaa_aug2001.pdf [last accessed: 11/07/2013].
- Thomas, G., 2009. *How to do Your Research Project*. London: Sage.

- Thomlinson, M.M., Robinson, M. and Challis, N.V., 2010a. Coursework, what should be its nature and assessment weight? *In: M. Robinson, N. Challis and M. Thomlinson, eds. Maths at University: Reflections on experience, practice and provision.* Birmingham, U.K.: More Maths Grads, pp. 122-126.
- Thomlinson, M.M., Robinson, M. and Challis, N.V., 2010b. Assessment and Retention. *In: M. Robinson, N. Challis and M. Thomlinson, eds. Maths at University: Reflections on experience, practice and provision.* Birmingham, U.K.: More Maths Grads, pp. 127-134.
- Tyrrell, S.E., 2002. Reflections on the evolution of a computer aided assessment strategy for statistics service teaching. *Maths-CAA Series*, June. Available via: www.mathstore.ac.uk/repository/mathscaa_jun2002.pdf [last accessed: 11/07/2013].
- University of East Anglia, 2012. *The MU-MAP Project*. Available via: www.uea.ac.uk/edu/mumap/ [last accessed 11/07/2013].
- Waldock, J., 2011. Learning, Teaching and Assessment Approaches to Developing Graduate Skills for Employability. *In: J. Waldock, ed. Developing Graduate Skills in HE Mathematics Programmes – Case Studies of Successful Practice.* Birmingham: Maths, Stats and OR Network, pp. 26-27
- Woodall, D.R., 2006. *LaTeXMathML: translating LaTeX math notation dynamically to Presentation MathML*. Available via: www.maths.nottingham.ac.uk/personal/drw/lm.html [last accessed 11/07/2013].
- Zheng, C.Y., 2011. Re: How to choose between Pearson and Spearman correlation? *Cross Validated*, 9 March. Available via: stats.stackexchange.com/a/8073 [last accessed 11/07/2013].

Appendices

A. List of relevant publications

The following parts of this thesis have been included in publications (presented in chapter order).

- The reflection on previous experience given in section 2.2 was published in the *Journal of Humanistic Mathematics*.
 - Rowlett, P.J., 2013. Developing a Healthy Scepticism About Technology in Mathematics Teaching. *Journal of Humanistic Mathematics*, 3(1), pp. 136-149. DOI: 10.5642/jhummath.201301.11.
- The literature review detailed in section 2.4 and survey of lecturers' views detailed in chapter 3 were submitted to the *International Journal of E-Assessment* in January 2013 as 'Mathematics lecturers' views on the advantages and limitations of e-assessment'.
- A brief account of the survey of lecturers' views detailed in chapter 3 was published in the proceedings of a Nottingham Trent University teaching and learning conference.
 - Rowlett, P., 2011. Mathematics lecturers' views of the advantages and disadvantages of electronic and traditional assessment. *Proceedings of Annual Learning and Teaching Conference "Learning and Technology"*, Nottingham Trent University, 30 March 2011. Nottingham, U.K.: Nottingham Trent University, pp. 19-20.
- Accounts of my previous experience of graduate skills development, outlined in section 5.3, were published in the proceedings of a Nottingham Trent University teaching and learning conference and a National HE STEM Programme report.
 - Rowlett, P., 2010. Guarding art galleries: design and evaluation of an undergraduate research based group project in mathematics. *Proceedings*

of Annual Learning and Teaching Conference "Developing students as scholars", Nottingham Trent University, 29 March 2010. Nottingham, U.K.: Nottingham Trent University, pp. 12-14.

- Rowlett, P., 2011. Using Art Gallery Problems to develop mathematical and employability skills in a higher education group project. *In: J. Waldock, ed. Developing Graduate Skills in HE Mathematics Programmes - Case Studies of Successful Practice. Birmingham, U.K.: Maths, Stats and OR Network, pp. 30-31.*
- The prototyping of a LaTeXMathML-based approach to processing MathML content and its trial in a blogging system, described in section 7.2.1, was included in an article in *MSOR Connections*.
 - Rowlett, P.J., 2008. Some approaches to mathematical blogging. *MSOR Connections*, 8(1), pp. 31-33. DOI: 10.11120/msor.2008.08010031.
- An account of the implementation of the partially-automated approach to assess individual work alongside a group project given in chapter 7 was submitted to the proceedings of the Nottingham Trent University Annual Learning and Teaching Conference 2013 in April 2013 as 'Assessing individual work alongside a group project - a partially-automated approach'.
- An abstract for a research paper based on the implementation and evaluation of the main experiment of this thesis, which will include elements of chapters 4, 5, 6, 7, 8 and 9, has been accepted for the proceedings of the British Congress of Mathematics Education (BCME) 8 conference as 'Development and evaluation of a partially-automated approach to the assessment of mathematics in higher education'.

B. List of relevant presentations

Seminars and conference talks

The following parts of this thesis have been presented in seminars and conference talks (presented in chapter order).

- The literature review detailed in section 2.4 was presented at a Nottingham Trent University teaching and learning conference.
 - Advantages and disadvantages of e-assessment and a partially-automated approach. *College of Arts and Science 2nd Annual Learning and Teaching Conference*, Nottingham Trent University (18/01/2013).
- The survey of lecturers' views detailed in chapter 3 was presented at a Nottingham Trent University teaching and learning conference and a mathematics department seminar at Newcastle University.
 - Mathematics lecturers' views of the advantages and disadvantages of electronic and traditional assessment. *Annual Learning and Teaching Conference 'Learning and Technology'*, Nottingham Trent University (30/03/2011).
 - Mathematics lecturers' views of e-assessment. *School of Maths and Stats seminar*, Newcastle University (21/03/2012).
- The discussion of qualities of assessment in chapter 4 and discovery of the approach to be taken detailed in chapter 5 were presented in a research seminar given to the School of Education at Nottingham Trent University.
 - A PhD project on e-assessment in mathematics: finding my approach. *School of Education Technology Enhanced Learning & Teaching Research Seminar Series*, Nottingham Trent University (15/05/2013).
- The approach to be taken detailed in chapter 5 was presented at the National HE STEM Programme Ideas Exchange 2012 conference.
 - An electronic hybrid approach to assessment. *Ideas Exchange 2012*, University of Sheffield (15/07/2012).
- The detail of previous experience of graduate skills development in section 5.3

was presented at: an HEA CETL-MSOR conference; a British Society for Research into Learning Mathematics conference; a Nottingham Trent University teaching and learning conference; National HE STEM Programme workshops at Sheffield Hallam University, University of York, Manchester Metropolitan University, University of Bath, University of Greenwich and University of Birmingham; and, a mathematics department seminar at Queen Mary, University of London.

- Group projects based on Art Gallery Problems. *CETL-MSOR Conference 2009*, Open University (08/09/2009).
- Using Art Gallery Problems to teach mathematical and employability skills in a higher education group project. *British Society for Research into Learning Mathematics Conference Autumn 2009*, Loughborough University (14/11/2009).
- Guarding Art Galleries: Design and evaluation of an undergraduate research based group project in mathematics. *Annual Learning and Teaching Conference 'Developing students as scholars'*, Nottingham Trent University (29/03/2010).
- Using Art Gallery Problems to Develop Mathematical and Employability Skills in a Higher Education Group Project. *Embedding Graduate Skills Workshop*, Sheffield Hallam University, (24/11/2010).
- Group projects based on Art Gallery Problems. *'Developing Graduate Skills in HE Mathematics Programmes' Workshop*, University of York (18/04/2011).
- Group projects based on Art Gallery Problems. *'Developing Graduate Skills in HE Mathematics Programmes' Workshop*, Manchester Metropolitan University (05/05/2011).
- Group projects based on Art Gallery Problems. *'Developing Graduate Skills in HE Mathematics Programmes' Workshop*, University of Bath (10/05/2011).
- Group projects based on Art Gallery Problems. *'Developing Graduate Skills in HE Mathematics Programmes' Workshop*, University of Greenwich

(17/05/2011).

- Group projects based on Art Gallery Problems. *'Developing Graduate Skills in HE Mathematics Programmes' Workshop*, University of Birmingham (19/05/2011).
- Graduate skills and interactive lectures. *School of Mathematical Sciences seminar*, Queen Mary, University of London (27/04/2012).
- The detail of the implementation of the partially-automated approach to assess individual work alongside a group project given in section 7.2 was presented at a Nottingham Trent University teaching and learning conference.
 - Assessing individual work alongside a group project – a partially-automated approach. *Annual Learning and Teaching Conference 'Student Transitions'*, Nottingham Trent University (26/03/2013).
- The evaluation of the partially-automated approach in sections 6, 7.3, 8 and 9 was presented at a Higher Education Academy conference.
 - Partially-automated approach to assessment of individual work alongside a group project in mathematics. *HEA STEM: Annual Learning and Teaching Conference 'Where practice and pedagogy meet'*, University of Birmingham (17/04/2013).
- An abstract for a talk about the design and delivery of the module described in chapter 5, which will include elements of the delivery and evaluation given in chapters 6, 7, 8 and 9, has been accepted for the CETL-MSOR Conference 2013 (10/09/2013) as 'Design and delivery of a new graduate skills module'.
- An abstract based on the implementation and evaluation of the main experiment of this thesis, which will include elements of chapters 4, 5, 6, 7, 8 and 9, has been accepted for the British Congress of Mathematics Education (BCME) 8 conference (14-17/04/2013) as 'Development and evaluation of a partially-automated approach to the assessment of mathematics in higher education'.

Non-live presentations

The outline findings of the survey presented in chapter 3 and a demonstration of the prototype implementation presented in section 7.2.1 were given in a short video on YouTube, 'Advantages and disadvantages of e-assessment', available from 25/09/2011. At the time of writing (17/07/2013) the website has recorded 453 views.

1. Advantages and disadvantages of e-assessment, 2011. *YouTube* - *peterrowlett*. 25/09/2011. Available via: www.youtube.com/watch?v=0JAeJKW7Hdc [last accessed: 12/07/2013].

C. Information given to second-markers of individual assignment 2

Grade descriptions (marking-notes.pdf)

As a general rule, the following applies (approximate classifications). However, individual pieces of work are moderated by complicated factors so these descriptions may not match every single piece of work (for example, a piece that gave the wrong answer in 1a) but considered the reality in 1b) very well might fall between these descriptions). Marks were also available for presentation which includes layout, use of connecting text, spelling, grammar, clarity, etc.

70+	In question 1: triangulates the polygon in part a); either in part a) or part b), either observes that the solution is minimal and discusses how to verify this or notices this is sufficient but not necessary; discusses the method used (particularly whether the triangulation is unique); uses an alternative mathematical method to produce an optimal solution; considers the reality in part b) including shifts, breaks, days off, annual leave, sickness and other statutory leave, hours worked per week, etc.; gives a clear description of the plan of work or a clear explanation of why the proposed number would not work.	In question 2: Gives a correct solution and shows some evidence of having read the theory.
60-70	In question 1: triangulates the polygon in part a); either in part a) or part b), either observes that the solution is minimal or observes that it is not; either finds a solution using fewer guards by another mathematical method which is still not optimal or finds an optimal solution by inspection only; considers the reality in part b) including most of shifts, breaks, days off, annual leave, sickness and other statutory leave, hours worked per week, etc.; gives a clear description of the plan of work or a clear explanation of why the proposed number would not work.	In question 2: Gives a correct solution.
50-60	In question 1: triangulates the polygon in part a); either in part a) or part b); perhaps reduces the solution by inspection only; considers the reality in part b) including some of shifts, breaks, days off, annual leave, sickness and other statutory leave, hours worked per week, etc.; gives a less clear description of the plan of work or explanation of why the proposed number would not work; perhaps identifies the unrealistic assumptions made but without taking account of them.	In question 2: Gives a solution which is correct using triangulation but actually fewer guards can be used.

40-50	In question 1: triangulates the polygon in part a); considers some aspects of reality in part b) but ultimately provides an unrealistic solution.	In question 2: Gives a solution which is correct using triangulation but actually fewer guards can be used.
-------	---	---

Mark scheme (marking-scheme.pdf)

Q1:

a) The answer given in the answers.pdf file is the answer.

b) This is hard to automatically provide an answer for as it is based on a free-text argument.

There is a temptation to think that 3 times the answer from a) is sufficient, since this would allow three shifts of enough workers to guard the museum. However, this does not take account of breaks, leave, illness, etc. or of legal limits on working hours. The student is presented with *approximately* 3 times the answer from a) and asked to say whether this is sufficient. Most marks are available for the argument that is made and the extent to which this relates the model to reality.

Marks:

- Answer to part a):
 - **1** for giving an answer which is either necessary or sufficient (i.e. three colouring the polygon to find $\left\lceil \frac{n}{3} \right\rceil$ guards for a polygon of n vertices gives an answer which is sufficient though not always necessary – this would attract this 1 mark).
 - **1** for using triangulation or quadrangulation.
 - **1** for making reference to mathematical theorems.
- Considering the method:
 - **2** for considering whether the triangulation or quadrangulation is unique.
 - **1** for considering the method otherwise, for example whether you can always triangulate.
- Finding the optimal solution:
 - **4** for noticing the solution is optimal, or for reducing it by another method (mathematical or inspection).
 - **2** for doing this reduction by a mathematical method (rather than “by inspection it looks like we can remove one guard”).

- In part b) use the number of guards proposed in part a) (follow-through).
- Answer to part b):
 - **1** for giving a clear answer yes or no.
- Reality:
 - up to **8** marks for the extent to which reality is considered, for example considering shifts, breaks, days off in a week, annual leave, sickness, hours of work per week.
- Explanation:
 - Score out of **5** for the clarity, reasonableness, level of detail, etc. of either the explanation of why not (if “no”) or the scheme of work that would make this work (if “yes”).

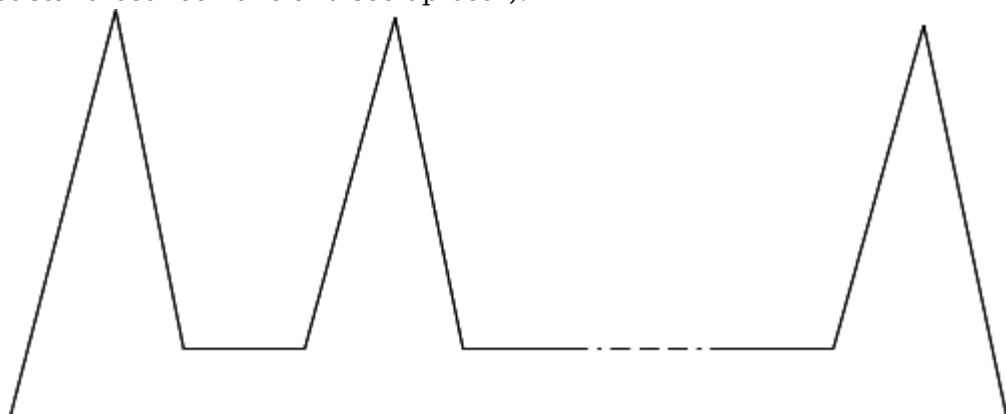
For example, an answer of “yes” to part b) might attract marks for a clear yes and for the scheme of work, but would probably lose the marks for reality. However, if part a) has been answered incorrectly, the answer to part b) might reasonably be yes.

Q2:

This is hard to automatically provide an answer for as it is based on a student-generated example.

There is a theorem that says that $\left\lceil \frac{n}{3} \right\rceil$ guards are sufficient for a polygon with n vertices, but not always necessary. Students may simply draw a n -sided polygon, claim $\left\lceil \frac{n}{3} \right\rceil$ and leave it at that, but such a polygon may not need that many guards unless it has been cleverly designed.

There is a canonical example for which $\left\lceil \frac{n}{3} \right\rceil$ guards are necessary given in the literature that looks like this (basically because every prong needs a guard as a guard cannot stand between two and see up both):



Though that is of course not the only form that works.

Marks:

- Correct answer:
 - Either **1** mark if the solution is the correct answer by triangulation only, or **3** marks if it is actually correct.
 - **2** marks for evidence of having read the literature on the topic: i.e. a museum like the one above would attract two marks as this is a basic example widely used in the literature; a variant of this approach might attract 1. If they have made their own with no reference to this sort of thing they do not get these two marks.

Both questions:

- A mark out of **8** for presentation, including for layout, use of connecting text, spelling, grammar, clarity, etc.

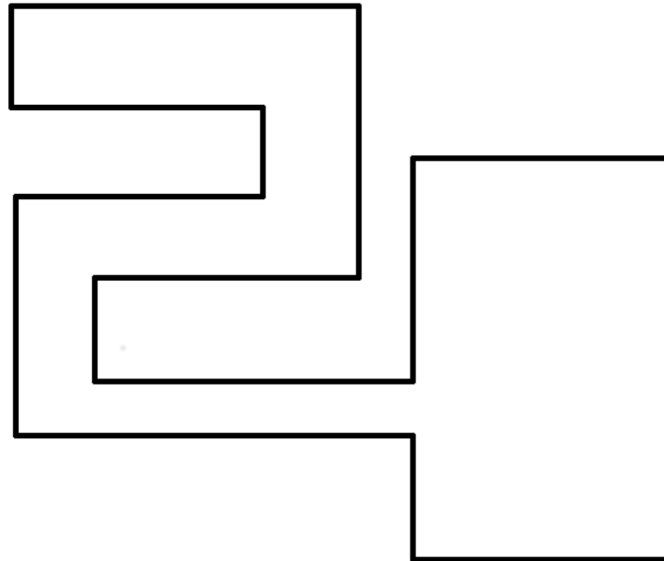
Sample piece of marked work

Question sheet

ID: 1

MAT30029: Individual assignment 2

1.



a) Show, by triangulating and three-colouring the polygon, how many guards are necessary to guard every point in the museum shown above at any one time.

b) In reality, given a staff of 11 guards, could you arrange for every point in the museum shown to be guarded 24 hours a day and seven days a week?

If so, how would you arrange this?

If not, why not and how many staff would you require?

2. Draw a polygon using 16 vertices for which 5 guards are necessary to guard every point at any one time.

Answer sheet

ID: 1

MAT30029: Individual assignment 2

1.

a) 3 guards (variant 1).

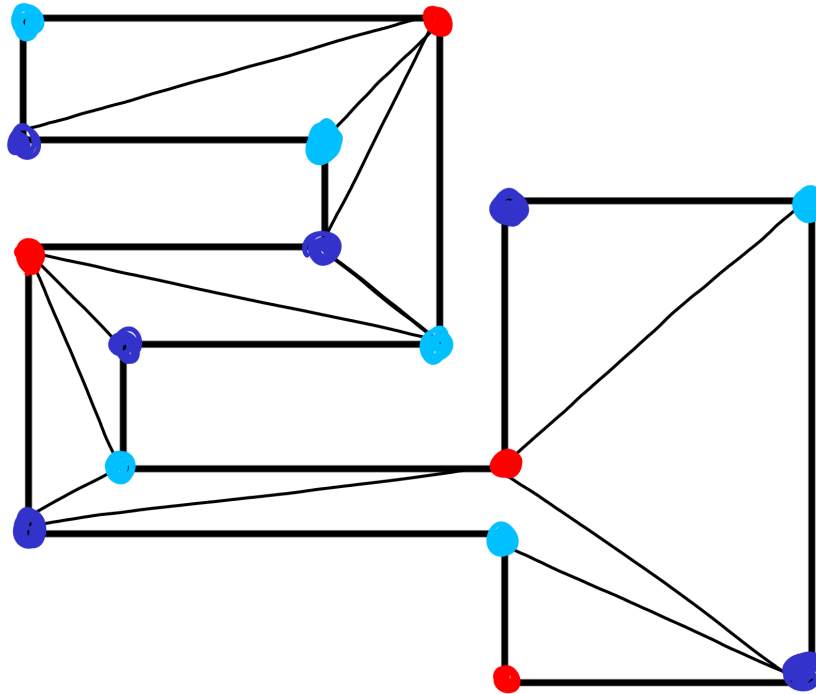
b) asking about 11 guards on staff.

2.

using 16 vertices for which 5 guards are necessary

Sample answer

Sample answer



1 a) The answer is three guards. ✓

The diagram has been triangulated (i.e. divided into triangles) and three-coloured. A theorem tells us that $\left\lceil \frac{n}{3} \right\rceil$ are sufficient for a polygon with n vertices, and this polygon has 16 vertices so 5 are sufficient, but not in fact necessary. Here I found the corridor to the right by considering multiple different triangulations (a theorem tells us you can always find a triangulation but not necessarily a unique one), searching for one which minimised the red dots to two. Two guards are indicated in the large space to the right, but this leaves two guards in essentially a six-sided convex polygon (albeit one that looks like it is arranged as a rectangle), and there is a theorem which says that any convex polygon only requires one guard. Choosing the guard by the entrance to the space allows us to cover the opening of the corridor as well. So there are four red guards and we remove one to leave three. ✓

b) Eleven guards are not enough to guard the museum. ✓

Nine guards could be arranged into three eight-hour shifts of three guards and therefore cover a 24 hour period. It is tempting to suggest that this leaves two to take account of days off. However, workers require breaks in an eight-hour shift, so each shift needs at least four members. Then, workers should be given one or two days off per week, an entitlement to

Answer - (1)
Triangulation - (1)
Ref. Theorem - (1)
Notice not optimal (4)
Uniqueness - (1)
Considers method - (1)

Reduction via
more mathematical
method than
simply inspection - (2)
Q1(a): 12/12

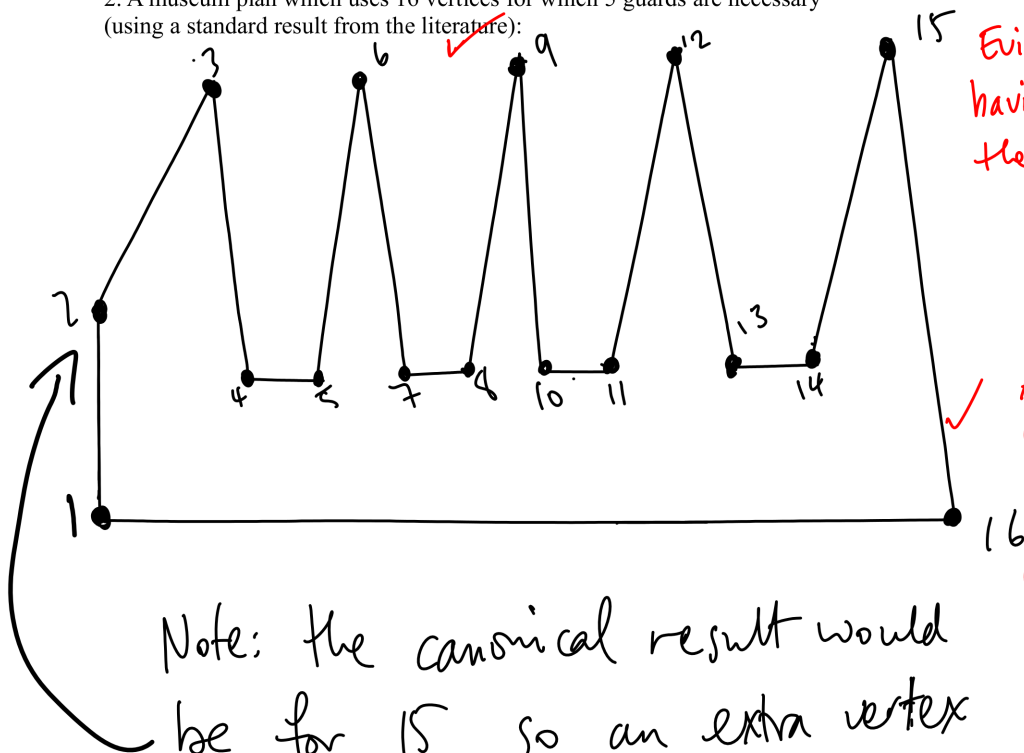
Answer (1)
Q1(b) Answer: 1/1

Days off
Breaks.

annual leave and allowance for sickness should be taken into account. Also, account should be taken of the number of hours worked in a week. Alternatively, it is possible to suggest that two twelve-hour shifts of four guards would provide 24 hour coverage, with the remaining three guards used to provide days off, etc. However, one day off per guard would mean they are working 72 hours a week, which contravenes relevant EU legislation. So practically and legally it is not possible to guard this museum 24/7 using a staff of eleven guards.

How many are actually required depends on the assumptions made and the preferences of the client. For example, a staff of seventeen guards could provide fifteen guards to work five days per week in three shifts and two guards to move between shifts, allowing four guards per shift each day with each guard working 40 hours per week and getting two days off per week. This would still not take care of annual leave, sickness, etc., for which an overtime scheme or additional (perhaps casual) staff members might be proposed.

2. A museum plan which uses 16 vertices for which 5 guards are necessary (using a standard result from the literature):



Note: the canonical result would be for 15, so an extra vertex is needed to make 16.

(Since $\lfloor \frac{16}{3} \rfloor = 5$.)

Marks/8 also for presentation (see mark scheme).

Annual leave.
Sickness.

Hours in working week.

Considers legislation.
Q1(b) Reality: 8/8

(Explanation should be scored /6 for clarity, how reasonable, etc.)

Evidence of having read the literature.
(2)

Actually correct!
(3)

Q2: 5/5